

Learning and Analytics in Intelligent Systems 15

S. Jyothi · D. M. Mamatha ·  
Suresh Chandra Satapathy ·  
K. Srujan Raju ·  
Margarita N. Favorskaya *Editors*

# Advances in Computational and Bio-Engineering

Proceeding of the International  
Conference on Computational and Bio  
Engineering, 2019, Volume 1

 Springer

# **Learning and Analytics in Intelligent Systems**

Volume 15

## **Series Editors**

George A. Tsihrintzis, University of Piraeus, Piraeus, Greece

Maria Virvou, University of Piraeus, Piraeus, Greece

Lakhmi C. Jain, Faculty of Engineering and Information Technology,  
Centre for Artificial Intelligence, University of Technology, Sydney, NSW,  
Australia;

KES International, Shoreham-by-Sea, UK;

Liverpool Hope University, Liverpool, UK

The main aim of the series is to make available a publication of books in hard copy form and soft copy form on all aspects of learning, analytics and advanced intelligent systems and related technologies. The mentioned disciplines are strongly related and complement one another significantly. Thus, the series encourages cross-fertilization highlighting research and knowledge of common interest. The series allows a unified/integrated approach to themes and topics in these scientific disciplines which will result in significant cross-fertilization and research dissemination. To maximize dissemination of research results and knowledge in these disciplines, the series publishes edited books, monographs, handbooks, textbooks and conference proceedings.

More information about this series at <http://www.springer.com/series/16172>

S. Jyothi · D. M. Mamatha ·  
Suresh Chandra Satapathy ·  
K. Srujan Raju · Margarita N. Favorskaya  
Editors

# Advances in Computational and Bio-Engineering

Proceeding of the International Conference  
on Computational and Bio Engineering, 2019,  
Volume 1

 Springer



*Editors*

S. Jyothi  
Department of Computer Science  
Sri Padmavati Mahila Visvavidyalayam  
(Women's University)  
Tirupati, Andhra Pradesh, India

D. M. Mamatha  
Department of BioScience and Sericulture  
Sri Padmavati Mahila Visvavidyalayam  
(Women's University)  
Tirupati, Andhra Pradesh, India

Suresh Chandra Satapathy  
School of Computer Engineering  
KIIT Deemed to be University  
Bhubaneswar, Odisha, India

K. Srujan Raju  
Department of Computer Science  
CMR Technical Campus  
Hyderabad, Telangana, India

Margarita N. Favorskaya  
Department of Informatics and Computer  
Techniques  
Siberian State Aerospace University  
Krasnoyarsk, Russia

ISSN 2662-3447                      ISSN 2662-3455 (electronic)  
Learning and Analytics in Intelligent Systems  
ISBN 978-3-030-46938-2              ISBN 978-3-030-46939-9 (eBook)  
<https://doi.org/10.1007/978-3-030-46939-9>

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Contents

<b>Polycystic Ovarian Follicles Segmentation Using GA</b> . . . . .	1
K. Himabindu, S. Narasimhulu, Ch. LawrenceDhreeraj, and T. Sarath	
<b>An Evolutionary Optimization Methodology for Analyzing Breast Cancer Gene Sequences Using MSAPSO and MSADE</b> . . . . .	9
K. Lohitha Lakshmi, P. Bhargavi, and S. Jyothi	
<b>Performing Image Compression and Decompression Using Matrix Substitution Technique</b> . . . . .	25
T. Naga Lakshmi and S. Jyothi	
<b>Classification of Cotton Crop Pests Using Big Data Analytics</b> . . . . .	37
R. P. L. Durgabai, P. Bhargavi, and S. Jyothi	
<b>Effect of Formulation Variables on Optimization of Gastroretentive In Situ Rafts of Bosentan Monohydrate HCl by 3<sup>2</sup> Factorial Design</b> . . .	47
B. Sarada, G. Srividya, R. V. Suresh Kumar, M. Keerthana, and M. Vidyavathi	
<b>Performance Analysis of Apache Spark MLlib Clustering on Batch Data Stored in Cassandra</b> . . . . .	65
K. Anusha and K. UshaRani	
<b>A Study on Opinion of B.Sc. Nursing Students on Health Informatics and EMR as Part of Nursing Education</b> . . . . .	77
B. GangaBhavani	
<b>A Comprehensive Hybrid Ensemble Method with Feature Selection Techniques</b> . . . . .	85
G. Sujatha and K. Usha Rani	
<b>DNA Based Quick Response (QR) Code for Screening of Potential Parents for Evolving New Silkworm Races of High Productivity</b> . . . . .	99
K. Haripriya, D. M. Mamatha, S. Jyothi, and S. Vimala	

<b>Identification of Neighbourhood Cities Based on Landuse Bigdata Using K-Means and K-NN Algorithm</b> .....	111
S. VinilaKumari, P. Bhargavi, and S. Jyothi	
<b>Secure Data Transfer Through Whirlpool—A Miyaguchi-Preneel Mode</b> .....	127
Prasanna Mala Chelamkuri, E. G. Bhavya Reddy, and Annapurnaeswari Jonna	
<b>Frequent Item-Set Mining Using Lexicographical Sequential Tree Construction on Map Reduce Framework</b> .....	135
P. Venkateswara Rao, D. Srinivasa Rao, and V. Sucharita	
<b>Deep Learning Based Recommender System Using Sentiment Analysis to Reform Indian Education</b> .....	143
Jabeen Sultana, M. Usha Rani, and M. A. H. Farquad	
<b>An Analysis of In Vitro Antioxidant and Anti-inflammatory Activities of <i>Mucuna pruriens</i> (Leaves) and <i>Allium sativum</i> (Bulbs)</b> .....	151
Bysani Jagannatha Divya, Bukke Suman, Mallepogu Venkataswamy, Kalla Chandra Mouli, and Kedam Thyaga Raju	
<b>A Novel Algorithm for Quality Evaluation Metrics of Fused Live Video Frames</b> .....	165
K. Sai Prasad Reddy, K. Nagabhushan Raju, and D. Sailaja	
<b>Cyber Crime Investigation and Law</b> .....	175
N. B. Chandrakala	
<b>Real Time Recognition of Rashdriving and Alcohol Detection to Avoid Accidents and Drunken Driving</b> .....	185
S. Swarnalatha, T. Srilakshmi, and K. Thilak Kumar	
<b>XGBoost Classifier to Extract Asset Mapping Features</b> .....	195
K. Sree Divya, P. Bhargavi, and S. Jyothi	
<b>Land Site Image Classification Using Machine Learning Algorithms</b> .....	209
G. Nagalakshmi, T. Sarath, and S. Jyothi	
<b>Decades of Research and Advancements on Fabrication and Applications of Silk Fibroin Blended Hydrogels</b> .....	219
Sufia Sultana, D. M. Mamatha, and Syed Rahamathulla	
<b>Long Non-coding RNA for Plants Using Big Data Analytics—A Review</b> .....	233
P. Swathi, S. Jyothi, and A. Revathi	

<b>In Silico, In Vitro and In Vivo Anti-inflammatory and Analgesic Activity of Usnic Acid</b> .....	249
D. Sujatha, Ch. Hepsiba Rani, Shaheen Begum, Sunitha Sampathi, and Saurabh Shah	
<b>Herbal Tea Treatment of Oligomenorrhea Condition with Hibiscus Rosa-Sinensis and Carica Papaya</b> .....	263
G. Sreesha and D. Sai Prasanna	
<b>Distribution and Evidential Incidence of Oral Microflora Among Dental Caries Infected 3–19 Year Old in Allahabad, India—A Pilot Study</b> .....	275
T. Jesse Joel, S. Sandeep Singh, and P. W. Ramteke	
<b>Screening of Genetic Variance Based on CO-I Gene Analysis of Silkworm (<i>Bombyx mori</i>) Races</b> .....	287
S. Vimala, Sriramadasu Kalpana, EI-Sheikh A. EI-Syed, and D. M. Mamatha	
<b>A Collaborative Filtering Based Ranking Algorithm for Classifying and Ranking NEWS TOPICS Using Factors of Social Media</b> .....	299
S. Gayathri Devi, K. R. Manjula, and K. Subhashri	
<b>Diversity Among Finger Millet Accessions Based on Genotyping Potential of SSR, EST-SSR and ISSR Markers</b> .....	319
Bheema Lingewara Reddy Inja Naga and S. Sivaramakrishnan	
<b>Social Media—Impact on Sexual and Reproductive Knowledge of Adolescents in South India</b> .....	335
N. Rajani and A. Akhila	
<b>Wearable Electronic Gloves in Two-Way Communication to Convert Signs into Speech</b> .....	345
S. Swarnalatha, Anusha Manubrolu, and Pooja Dande	
<b>A Perspective Overview on Machine Learning Algorithms</b> .....	353
S. Nalini Durga and K. Usha Rani	
<b>A Methodology for Detecting ASD from Facial Images Efficiently Using Artificial Neural Networks</b> .....	365
T. Lakshmi Praveena and N. V. Muthu Lakshmi	
<b>Service Composition in Mobile Ad Hoc Networks (MANET's) with the Help of Optimal QoS Constraints</b> .....	375
G. Manoranjan, M. V. Rathnamma, V. Venkata Ramana, and G. R. Anil	
<b>Inferential Procedures for Testing Assumptions on Observations for Applications of Biometric Techniques</b> .....	391
M. Naresh, B. Sarojamma, P. Srivyshnavi, G. Madhusudan, P. Vishnupriya, and P. Balasiddamuni	

<b>Smart Crop Suggester</b> .....	401
N. Usha Rani and G. Gowthami	
<b>The Role of Long Non-Coding RNA (lncRNA) in Health Care Using Big Data Analytics</b> .....	415
A. Revathi, S. Jyothi, and P. Swathi	
<b>A Framework for Modeling and Analysing Big Biological Sequences</b> .....	429
Sai Jyothi Bolla and S. Jyothi	
<b>Specification and Estimation of a Biometric Model by Using Logistic Regression for Measuring Child Mortality</b> .....	439
P. Vishnu Priya, B. Sarojamma, G. Madhusudan, P. Srivyshnavi, M. Naresh, and P. Balasiddamuni	
<b>A Case Study Report: Ruptured Scar Ectopic Pregnancy</b> .....	447
Abhilaasha Macherla and R. V. Raviteja	
<b>Some Modified Biometrical Diversity and Evenness Indices</b> .....	451
G. Madhusudan, P. Srivyshnavi, B. Sarojamma, M. Naresh, R. Abbaiah, and P. Balasiddamuni	
<b>Data Analysis on Biopsies of Breast Cancer Tumors Data Using Data Science</b> .....	461
K. Hemalatha, K. Hema, and V. Deepika	
<b>A Comparison of Multi Support Vector Machine Performance with Popular Decomposition Strategies on Alzheimer's Data</b> .....	469
R. M. Mallika, K. Usha Rani, and K. Hemalatha	
<b>Synthesis, Evaluation and in Silico Studies of 4-N, N-Dimethylamino and 4-Carboxy Chalcones as Promising Antinociceptive Agents</b> .....	481
Shaheen Begum, S. K. Arifa Begum, A. Mallika, and K. Bharathi	
<b>In Silico Analysis for Detection of Glucose Transport-2 Inhibitors from Seagrass</b> .....	491
Mathakala Vani, Narem Ritesh Siddhartha Reddy, and Palempalli Uma Maheswari Devi	
<b>Automated Diagnosis of Shoulder Pain Using Regression Algorithms</b> .....	499
B. Triveni, P. Bhargavi, and S. Jyothi	
<b>Diagnosis of Urological Diseases Using Deep ROI</b> .....	515
R. Venkata Raviteja, M. Abhilaasha, and B. Prakasha Rao	

<b>Pharmacokinetic and Pharmacodynamic Studies on Celecoxib Loaded Nanosponges Gel for Topical Delivery</b> . . . . .	525
Y. Sarah Sujitha and Y. Indira Muzib	
<b>Smart Bed Companion</b> . . . . .	545
G. V. V. S. Naveen, M. Shivani, Jalla Hasmitha, and D. Ajitha	
<b>Onion Husk Powder as a Adsorbent for Removal of Methylene Blue and Malachite Green from Aqueous Solutions</b> . . . . .	549
R. Usha, Ch. Indhravathi, D. Hymavathi, and M. Vijayalakshmi	
<b>Bioalgalization—A Novel Approach for Soil Amendment to Improve Fertility</b> . . . . .	557
Layam Anitha, Gannavarapu Sai Bramari, and Pilla Kalpana	
<b>GC-MS Analysis and Computational Studies of Roots of Anthocephalus Cadamba</b> . . . . .	569
Kaveripakam Sai Sruthi and Adikay Sreedevi	
<b>Comparative Omics Based Approach to Identify Putative Immunogenic Proteins of <i>Trichomonas Foetus</i></b> . . . . .	583
Geethanjali Karli, Rathnagiri Polava, and Kalarani Varada	
<b>SVM Based Approach to Text Description from Video Sceneries</b> . . . . .	593
Ramesh M. Kagalkar, Prasad Khot, Rudraneel Bhaumik, Sanket Potdar, and Danish Maruf	
<b>Social Networking a Peril to Youth and Cultural Nuances—Needs Legal Fortification</b> . . . . .	601
G. Indirapriyadarsini and P. Neeraja	
<b>A Statistical Study on Analysing Repeated Measures of Data of Hyperlipidemia Cases</b> . . . . .	613
M. Siva Parvathi, K. Blessy Deborah, R. Vishnu Vardhan, T. Sukeerthi, and K. Sukanya	
<b>Integrated Geospatial Technologies in e-Governance: An Indian Scenario</b> . . . . .	633
Pondari Satyanarayana, S. Jyothi, and Dandabathula Giribabu	
<b>Molecular Properties Prediction of N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-Substituted Phenylacrylohydrazides</b> . . . . .	641
K. Saritha and G. Rajitha	
<b>Chitosan as a Heavy Metal Adsorbent in Waste Water Treatment</b> . . . . .	649
M. Saraswathi and R. J. Madhuri	

**Perceptions of Adolescents on Hazards of Using Electronic Gadgets** ..... 655  
D. Jyothi and E. Manjuvani

**Genomics in Big Data Bioinformatics**..... 661  
Tahmeena Fatima and S. Jyothi

# Polycystic Ovarian Follicles Segmentation Using GA



K. Himabindu, S. Narasimhulu, Ch. LawrenceDhreeraj, and T. Sarath

**Abstract** Up to 5–15% of the women affects the reproductive system this abnormality syndrome called Polycystic Ovarian Syndrome (PCOS). Polycystic ovary syndrome (PCOS) has been a gynecological endocrine syndrome that proffers the consequence in health issues of menstrual dysfunctions, androgynism and also infertility. Usually it occurs in reproductive aging women. PCOS directs to unsuitable follicle development of the ovaries that are seized at a former stage. Periodic measurements of the dimension and description of follicles over several days are the crucial means of enquiry by physicians. In this paper, a new algorithm for automatic detection of follicles in ultrasound image for ovaries is suggested. The proposed algorithm uses various edge based methods are using for Ovaries follicles segmentation that is GA with Sobel and GA with Canny. Hence, we compare the variety of these techniques and demands assures the GA with Canny operator provides a better performance on ovarian follicle.

**Keywords** Ovarian follicle segmentation · Genetic Algorithm · Edge based methods · Polycystic ovary syndrome

---

K. Himabindu (✉) · Ch. LawrenceDhreeraj  
Department of MCA, SVCE, Tirupati, India  
e-mail: [bindukar.pujari@gmail.com](mailto:bindukar.pujari@gmail.com)

Ch. LawrenceDhreeraj  
e-mail: [dheerajchakram@gmail.com](mailto:dheerajchakram@gmail.com)

S. Narasimhulu  
Department of CSE, SVCE, Tirupati, India  
e-mail: [narasimhulu.sangi@gmail.com](mailto:narasimhulu.sangi@gmail.com)

T. Sarath  
Department of CSE, SISTK, Puttur, India  
e-mail: [sarathbalakrishna@gmail.com](mailto:sarathbalakrishna@gmail.com)

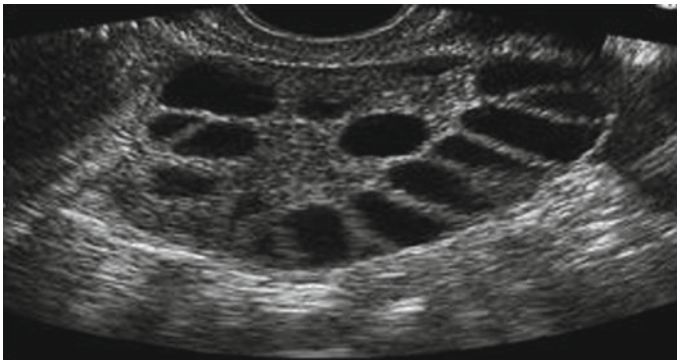
© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_1](https://doi.org/10.1007/978-3-030-46939-9_1)



## 1 Introduction

Polycystic ovary syndrome (PCOS) could be an intricate state characterized by elevated hormone levels, discharge irregularities, and/or tiny cysts on one or each vary [1]. Generally, up to 5–14% women at the age of 18–44 years are suffering from Ovarian Syndrome, it makes most common endocrine irregularity among reproductive age [2]. Hence women need to consult health care professionals to resolve problems with fat, acne, amenorrhea, extreme hair growth, and infertility usually receive a verdict of PCOS. Most of the women with PCOS are suffers with endovascular cancer, cardiovascular disease, dyslipidemia, and type-2 diabetes [3]. The ovarian follicles are having structured with filled spherical fluid. Polycystic ovarian syndrome ultrasound image shown in Fig. 1.

During the follicular phase, a tiny low cohort of follicles begins to develop. A whole perceptive of ovarian follicle dynamics is vital within the field of biotechnology and human reproduction. For ladies endure assist generative medical aid, the ovarian ultrasound imaging has turn into a valuable tool in infertility management. Periodic dimensions of the follicles size over many days the first suggests of analysis by the physician [4]. Polycystic Ovarian (PCO) ultrasound image is analyzed by the quantity of follicles, follicle size identification, that distribution, and evaluate the number of follicles ratio to ovarian volume. Detection of PCO, pelvic ultrasound image is important and gives accurate result. Most of the cases the analysis of ultrasound images are physically. So far there is so much of variance occurs among different gynecologists/radiologists. Hence, in this paper to detect the PCO stage segment the pelvic ultrasound imagery and to detect the edges in a PCOS image is an exigent task and to use a variety of edge detection algorithms on PCOS ultrasound imagery. Then it provides comparison description of edge detection on segmentation.



**Fig. 1** Polycystic ovary syndrome

## 2 Image Segmentation

Image segmentation is the method of segment a digital image into various segments [5]. To modify and/or remodel the image into meaningful for easy analysis, this is the main objective of segmentation. The segmentation have various techniques. It is sub divided the image into several parts, the aim of image segmentation that can be customized and analyzing easily [6]. Image segmentation is specially used to trace the object and boundaries. Each of the pixels in a region is related to some uniqueness or compute accurately. Currently we have a tendency to discuss concerning the conception of edge detection. Edge detection may be extremely developed within the image process. Region boundaries and edge are one among the techniques that are closely connected with edge detection via segmentation. The edges consist of many mathematical strategies to focus at coordinates in digital image segmentation and additionally image brightness with change accuracy are more professionally as discontinuous via various strategies of edge detection. It principally targeted on feature detection and feature extraction [7].

### 2.1 Edge Detection with Sobel Operator

The method of Sobel edge detection for image segmentation finds edges via the Sobel estimate to the derived. It precedes the edges at those points where the grade is maximum. The Sobel methods perform a 2-D spatial gradient compute on a picture so highlights region of high spatial frequency that communicate to edges. Generally, it's habituated notice and calculable absolute gradient magnitude at every purpose in n input gray scale image [8]. This methodology relies on convolving the image. The Sobel image has tiny, distinguishable and numeral valued filter, arise the horizontal and vertical track. Its low expansive within the term of computation. As per the Sobel operator PCOS ultrasound image shown in Fig. 2.

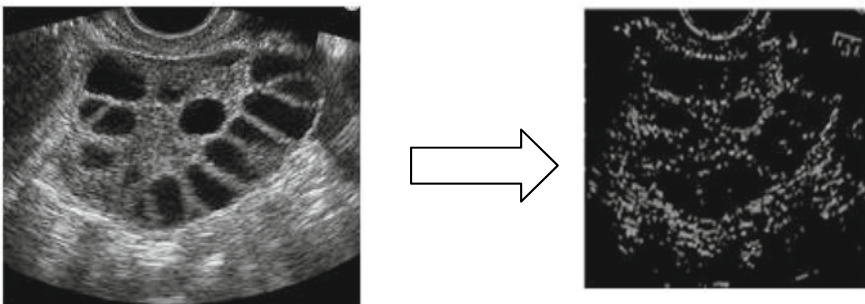
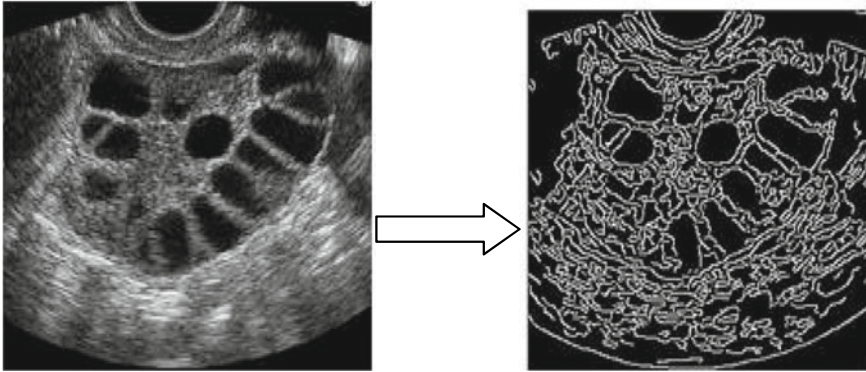


Fig. 2 Follicles segmentation with Sobel operator



**Fig. 3** Follicles segmentation with Canny operator

## ***2.2 Edge Detection with Canny Operator***

To find edges by separating noise from the image is completed by Canny edge detection that may be an important technique. Canny edge detection is a technique to extract helpful structural in sequence from dissimilar vision objects and significantly decrease the amount of information to be processed. Canny methodology is superior methods without worrying the options of the edges within the image later on it applies the tendency to find the edges and also the serious worth for threshold. Hence, an edge detection justifies to think about these necessities are often enforced in a very broad sort of the positions [9, 10] (Fig. 3).

## ***2.3 Genetic Algorithm for Edge Detection***

The follicle edges are formed based different edge detection operators are used for identifying the follicle size of PCOS ultrasound images. For improving the edges of ovaries this paper Sobel, Canny, Sobel with Genetic Algorithm and Canny with Genetic Algorithm edge detection operators are applied on ultrasound images for identifying the ovary size [11–14]. Generally, Genetic Algorithm (GA) is used for reducing the Mean Square Error (MSE). The GA algorithm follows under considerable steps [15].

Algorithm:

- Step 1 First select the edge detected input image of PCOS ultrasound image.
- Step 2 For 3 \* 3 operator mask applying Genetic Algorithm.
- Step 3 Then Perform above masking operators edge detection on the selected image.
- Step 4 Finally compare the result obtained image with ideal expected output image using on GA fitness function and update the mask.

Step 5 Repeat Step 3 until optimization gets stopping criteria. Step 6: finally, the result is shown.

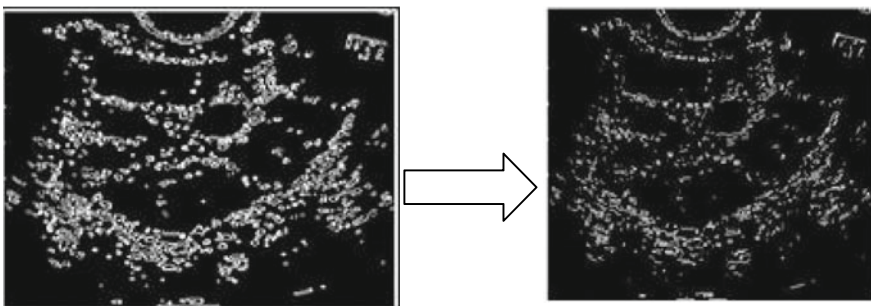
### 3 Result and Discussions

The combination of Sobel with GA optimized operator and Canny with GA optimized operator applied on PCS ultrasound image for getting better edge detection shown in Table 1 (Figs. 4 and 5).

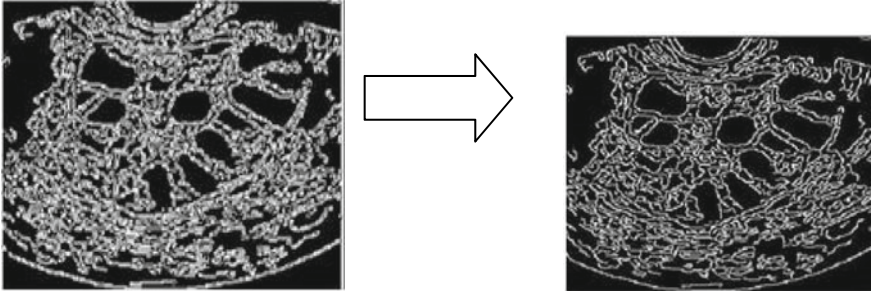
According to Table 2, GA optimization operator was applied on ultrasound PCOS image. And the table describes comparison of output before applying GA optimization and output after applying GA optimization. Here two performance parameters are used that is PSNR (Peak Signal to Noise Ratio), MSE (Mean Square Error). Higher PSNR is always provides the better image quality. Hence, in this work GA with Canny operator provides the better edge image comparing to GA with Sobel operator.

**Table 1** GA optimized operator with Sobel and Canny operators output for PCOS

Generation	f-count	Best f(x)	Max. constraint	Stall generations
1	10,400	0.599056	0.01535	0
2	20,600	0.592432	0.03143	0
3	30,800	0.58887	0.0009729	0
4	41,000	0.585385	0.04252	0
5	51,200	0.592975	0.002514	0
6	61,400	0.592975	0.002514	1
7	71,600	0.592786	0.592786	1
8	81,800	0.591543	0.0004562	0



**Fig. 4** Sobel with Genetic Algorithm



**Fig. 5** Canny with Genetic Algorithm

**Table 2** MSE and PSNR output value before and After GA optimization for PCOS image using Sobel and Canny operator

Calculated value	Output before GA optimization for PCOS image		Output after GA optimization for PCOS image	
	Sobel	Canny	Sobel	Canny
MSE	0.6062	0.6069	0.5915	0.5929
PSNR	5.0061	5.0073	5.2502	5.2512

$$PSNR = 10 \cdot \log_{10} \left( \frac{MAX P_I}{MSE_r} \right) \quad (1)$$

where MAX PI represents maximum pixel value in the image.

## 4 Conclusion

In this paper, for ovarian follicle segmentation we have applied different techniques such as Canny, Sobel, GA with Sobel and GA with Canny edge detection operators on PCOS ultrasound images. In this approach GA optimization was done before and after. Finally, after GA with Canny optimization operator provides the accurate result for PCOS image follicle segmentation.

## References

1. E.M. Umland, L.C. Weinstein, E.M. Buchanan, in *Menstruation-Related Disorders*, ed. by J.T. DiPiro, R.L. Talbert, G.C. Yee et al. Pharmacotherapy: A Pathophysiologic Approach, 8th edn. (McGraw-Hill, New York, 2011), p. 1393
2. L.H. Lin, M.C. Baracat, A.R. Gustavo et al., Androgen receptor gene polymorphism and polycystic ovary syndrome. *Int. J. Gynaecol. Obstet.* **120**, 115–118 (2013)

3. M. Aubuchon, R.S. Legro, Polycystic ovary syndrome: current infertility management. *Clin. Obstet. Gynecol.* **54**(4), 675–684 (2011)
4. N. Kita, I. Georgiou, A. Tsatsoulis, The genetic basis of polycystic ovary syndrome. *Eur. J. Endocrinol.* **147**, 717–725 (2002)
5. F.C. Monteiro, A. Campilho, A.: Watershed framework to region-based image segmentation. in *Proc. International Conference on Pattern Recognition, ICPR 19th*, pp. 1–4, (2008)
6. M. Hameed, M. Sharif, M. Raza, S.W. Haider, M. Iqbal, Framework for the comparison of classifiers for medical image segmentation with transform and moment based features. *Res. J. Recent Sci.* **2277**, 2502 (2012)
7. A. Fabijanska, Variance filter for edge detection and edge-based image segmentation, in *Proceedings International Conference on Perspective Technologies and Technique in MEMS Design (MEMSTECH)* (2011), pp. 151–154
8. V. Sucharita, S. Jyothi, D.M. Mamatha, A comparative study on various edge detection techniques used for the identification of penaeid prawn species. *Int. J. Comput. Appl.* **78**(6), 0975–8887 (2013)
9. N. Marina, T. Eva, T. Milan, Tuba, in *Edge Detection in Medical Ultrasound Images using Adjusted Canny Edge Detection Algorithm*. IEEE Xplore, Electronic ISBN: 978-1-5090-4086-5. <https://doi.org/10.1109/telfor.2016.7818878> (2017)
10. C. Panchasara, *Application of Image Segmentation Techniques on Medical Reports*. vol. 6, no. 7 (2015), pp. 2931–2933
11. K. Himabindu, S. Jyothi, D.M. Mamatha, *GA Based Feature Selection for Squid's Classification*, vol. 2 (2018). [https://doi.org/10.1007/978-981-13-3393-4\\_4](https://doi.org/10.1007/978-981-13-3393-4_4)
12. P. Mantas, U. Andruis, A survey of genetic algorithms applications for image enhancement and segmentation. *Inf. Technol. Control* **36**(3), 278–285 (2007)
13. F. Saitoh, Image contrast enhancement using genetic algorithm, in *IEEE International Conference on Systems, Man, and Cybernetics, IEEE SMC'99*, vol. 4 (1999), pp. 899–904
14. L. Caponetti, N. Abbattista, G. Carapella, A genetic approach to edge detection, in *IEEE International Conference on Image Processing*, vol. 1 (1994), pp. 318–322
15. M. Lee, K. Leung, S.W. Pun, T.L. Cheung, EDGE detection by genetic algorithm, in *Proceedings 2000 International Conference on IEEE Transactions on Image Processing*, vol. 1, pp. 478–80. (2000)

# An Evolutionary Optimization Methodology for Analyzing Breast Cancer Gene Sequences Using MSAPSO and MSADE



K. Lohitha Lakshmi, P. Bhargavi, and S. Jyothi

**Abstract** An evolutionary methodology using multiple sequence alignment technique with optimal search algorithms particle swarm optimization and differential evaluation is proposed in this paper. Proposed methodology algorithms are termed as Multiple Sequence Alignment and Particle Swarm Optimization (MSAPSO) and Multiple Sequence Alignment and Differential Evaluation (MSADE). These techniques are developed to categorize gene sequences based on optimal result produced for each generation. These evolutionary techniques encompasses of two phases in designing. In first phase MSA is pragmatic on pair wise sequences to generate aligned sequence as output. The sequence generated as output in the first level will be given as input to second phase. In the second segment optimal search algorithms PSO or DE are applied on sequences which generate optimal value and generation best value for each generation. These values are considered for further categorization. This paper presents analysis of MSAPSO and MSADE on gene sequences.

**Keywords** Multiple sequence alignment (MSA) · Particle swarm optimization (PSO) · Differential evaluation (DE) · Breast cancer sequence categorization

## 1 Introduction

There has remained a developing increment in the frequency of breast cancer mostly in women leads to cause female death [1, 2]. Despite huge advancement in progress of breast cancer, the scan for cause of incidence of disease and experiments on healing treatment is being conducted rapidly [3]. Heredity is also one cause for spread of this disease and genome-wide Association Studies (GWAS) have recognized that

---

K. Lohitha Lakshmi (✉) · P. Bhargavi · S. Jyothi  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [lohita.kanchi@gmail.com](mailto:lohita.kanchi@gmail.com)

P. Bhargavi  
e-mail: [pbhargavi18@yahoo.co.in](mailto:pbhargavi18@yahoo.co.in)

S. Jyothi  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_2](https://doi.org/10.1007/978-3-030-46939-9_2)



more than 180 normal hereditary variations related with breast cancer growth [4]. Only before it is assumed that only 5–10% of breast malignancy is caused due to heredity and is caused by transmissible mutations in considered suspected genes like BRCA1 and BRCA2 [5]. As gene sequencing technologies progress, as number of clinical trials have been lead on BRCA1 and BRCA2 genes it is considered that inherited mutations in these genes can cause breast cancer in 50% of patients. Some more genes have been revealed which become most important susceptible in cause of breast cancer in 20–30% patients such as BRIP1, PALB2, TP53, PTEN, STK11, CDH1 etc. [6]. Next generation sequencing proposals an innovative destination for the analysis of gene sequences which plays a vital role in risk assessment of genetic diseases. Before this evolution risk is predicted only by family history rather than genetic information by using standard models [7]. Bioinformatics can be well-defined as the solicitation of software computing techniques in the field of biological applications mainly in genomic related information which comprises of vast, inaccurate, incomplete, and ambiguous real time data. Soft computing is progressively opening up several techniques to produce possible and accurate optimal solutions in genetic research [8]. The aim of SC is to provide abilities for handling real life vague and incomplete data and have close similarity with human alike decision-making which is essential for genetic data analysis. Among SC constituents evolutionary algorithms play major role in genetic related experiments [9].

In this paper from evolutionary algorithms of SE mainly particle swarm optimization and differential evaluation techniques are applied with the combination of Multiple Sequence Alignment (MSA) on genetic data. MSA is used in detecting structural and also functional similarities of two or more sequences in order to reliably detect evolutionary associated sequences by discovering comparable positions among set of sequences [10]. These paired sequences are given as input to population centered stochastic search techniques PSO and DE which can be applied on genetic related experiments to get optimal solution. In my present work, these strategies are proposed to get optimal solution by applying on hereditary genetic diseases by applying medical data sequences interrelated to a particular disease identical breast cancer, leukemia, color blindness, diabetes etc. These evolutionary algorithms are applied on each trail population and change the population randomly and include selection processes to assess which solution is more adaptable to forthcoming generation scientific experiments [11].

## 2 Soft Computing Techniques for Gene Sequence Analysis

The main components of SC are Fuzzy Logic (FL), Evolutionary Algorithms (EA), Artificial Neural Networks (ANN), and Probabilistic Reasoning (PR). Among them evolutionary algorithms are more efficient to be applied on bioinformatics related



applications to get global optimization. SC computational approach has close resemblance to human mind to deal with unclear and indefinite data. Gene sequence analysis is used to identify similarities and transformations of different characteristics between organisms in a same species [12].

## ***2.1 Particle Swarm Optimization (PSO)***

PSO is an evolutionary optimization method presented in soft computing techniques. PSO is a population centered optimization procedure which finds similarity between populations belongs to a same species by which every move and position of the molecule leads to a solution to the problem. This phenomenon is applied on flocking birds. Every bird in the flock is measured as a molecule and its speed and position is considered respective to its past conduct (light of other bird) and of its own [13]. In instance of flocking birds this strategy works to find best hunt regions and in case of gene sequence analysis this works to find structural similarity between gene sequences belonging to a particular disease to analyses behavior of an individual characteristic's. Due to the capability of finding structural similarity PSO is producing optimal results in hereditary (genetic) diseases in factual time genetic experiments [14]. PSO is a population based iterative algorithm. For individual iteration, computation will be performed on individual population based on fitness function and best fittest individual will be nominated as a parent to that generation [15] simple PSO algorithm can solve only single objective optimization problems. To resolve multi objective optimization complications some extension to PSO is needed to find best solution [16].

## ***2.2 Differential Evolution (DE)***

DE is a procedure which creates vector differences by using an iterative mutation technique to produce new candidate solution for each generation. In most realistic experiments related to genetics, bioinformatics, and computational biological experiments attaining global optimization results are playing a vital role. A population matrix with optimized values will become the output of implementation of DE algorithm. Best individuals are selected as solution from this population with best optimal values. But in real time applications simple DA implementation is not sufficient due to its large number of functional assessments and unusual runtime complexity sometimes ranges from times to days [17]. In my earlier work I have implemented DE algorithm on different groups of gene sequences and considered Generation Best [GB] value for individual generation to compare computational accuracy to discriminate between various categories of sequences to categories breast cancer and

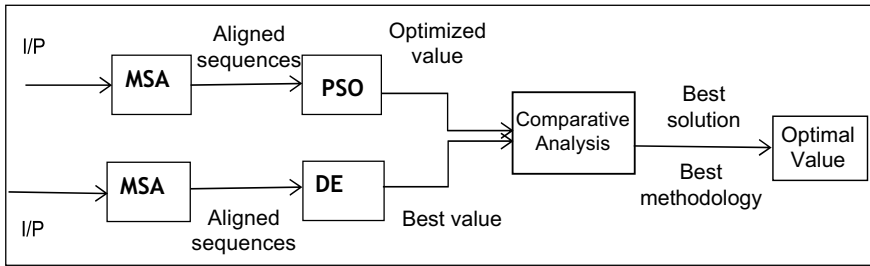
non-breast cancer sequences [18]. To enhance computational accuracy some extension is needed for simple DE implementation in real time applications on problem vector.

### 3 MSA for Gene Sequence Analysis

The main purpose of Multiple Sequence Alignment (MSA) is to align two or more than two sequences to find functional and structural similarities between gene sequences to predict the possibility of hereditary features, properties, or hereditary diseases to next generation or child generation. It is strongly suggested that more the sequence similarity there is greater chance of having similar structure which leads to have similarity functionality. MSA is a method comprises of sequential set of algorithms implemented on evolutionary sequences. MSA is a tool developed to address very complex biological computational problems like sequence analysis [19]. Next generation sequencing meant to deal with large volume of raw data sequences. In order to advance performance analysis of large scale data sets it is essential to combine MSA algorithm techniques with soft computing techniques to improve the speed, precision, and reliability in producing near optimal global solution. In this present work while implementing multiple sequence alignment mainly focused on algorithms like CLUSTALW (Cluster Analysis of Pairwise Alignment) and PRRN. Among many tools CLUSTALW and PRRN is more acceptable tool for present implementation. CLUSTALW is unique of the MSA algorithms which can combine global pairwise alignment in addition progressive method [20]. PRRN produces optimal MSA alignment score by using hill-climbing algorithm.

### 4 Proposed Evolutionary Methodology

Previous work presented experimental resultant values applied on cancer and cancer suppressor (non-cancer) gene data sequences with particle swarm optimization and differential Evaluation. In previous observed PSO algorithm implementation values on cancer and cancer suppressor sequences, The values are ranges approximately between 9000 and 15,000 for 95% of cancer sequences and the values ranges from 25,000 to 61,000 for 95% of cancer suppressor gene sequences. When final values are observed after implementation of PSO algorithm it is concluded that resultant optimal values generated after implementing PSO algorithm on breast cancer diseased input sequences are less than the breast cancer suppressor (normal) gene data input sequences [13]. Earlier this method paved a way to distinguish concerning cancer and non-cancer sequences based on generated optimal values. Subsequently based on PSO implementation I have proposed another methodology by Differential Evaluation algorithm on breast cancer besides normal breast gene data sequences. In this implementation different categories of sequences are distinguished based on



**Fig. 1** Architecture of proposed hybrid methodology

the DE algorithm implementation generated optimal values vice versa. When results are detected with both executions of PSO and DE algorithms individually on gene sequences optimal result values obtained with DE implementation is more accurate when compared to PSO implementation.

#### **4.1 Implementation of Evolutionary Methodologies MSAPSO and MSADE**

To advance the accuracy percentage of experimental values and to validate a certain category of classification, combination of some advanced methods should be added to existing algorithms [21]. Successive combination of different techniques leads to an evolutionary methodology to advance performance accurateness [22]. In the proposed evolutionary hybrid methodology Multiple Sequence Alignment (MSA) remains combined with both PSO and DE, which leads to propose evolutionary methodologies MSAPSO and MSADE to perform relative analysis of breast cancer and normal breast gene sequences to advance efficiency and worth solution with improved calculation accuracy to discriminate different sorts of sequences (Fig. 1).

#### **4.2 Algorithm for MSAPSO**

1. Read mRNA or DNA structures from NCBI site.
2. Start the MSA process with the prerequisite gene input sequences.
3. Align first sequence with second sequence which belongs to same category of sequences (i.e. genetic data sequences of particular disease and healthy sequences).
4. Distance matrix will be created of same category for every pair of sequences.
5. Optimal value will be produced based on alignment score by using Hill climbing algorithm built on the methodology (CLUSTALW or PRRN).

$$[f(X) = e - (x^2 + y^2)] \quad (1)$$

$f(X)$  is a target function, where  $(x, y)$  is a vector with constant or discrete values.  $f(X)$  accepts any change and iteration endures until no change found in generated or final vector.  $X$  is assumed to be local optimal a surface with only one maximum and this will be converged to global maxima.

6. Sequence generated relevant to optimal score will be measured as resultant sequence and this generated sequence will be passed as input to PSO algorithm.
7. Start PSO function by Replacing  $N, A, C, G,$  and  $T$  values by 0, 1, 2, 3, and 4 values.
8. Initialize parameters starting location. Velocity of the element and individual best position to zero and individual best error value and individual error value with  $-1$ .
9. Set the input bounds to some values which specify the beginning position of the particle movement and closing position of the particle movement.
10. Invoke PSO function by initializing best error value for group to  $-1$  and best position for the group to zero.
11. Establish swarm and begin optimization loop through particles in swarm. Evaluate the fitness value using required user defined mathematical equation based cost function.
12. Check the current position ( $x_i$ ) of the particle to determine best or not. There are three unique forces working on each particle. They are particles initial velocity ( $v_i$ ), position at time step  $t$  and distance from the individual particle that is cognitive force and separation from the swarm's best known position called social force [21].

$$[x_i(t + 1) = x_i(t) + v_i(t + 1)] \quad (2)$$

13. Update particle position and velocities for each particle through swarm.

$$\begin{aligned} V_{i,j}(t + 1) = & V_{i,j}(t) + c1r1.j(t)[v_{i,j}(t) - x_{i,j}(t)] \\ & + c2r2.j(t)[y'_{i,j}(t) - x_{i,j}(t)] \end{aligned} \quad (3)$$

$r1.j(t), r2.j(t) \sim U(0, 1)$  are uniform random numbers in the range  $[0, 1]$   
 $y'_{i,j}(t)$ —position vector of neighbor's best particle

14. Steps from 6 to 8 are rehashed until the best position (globally) is produced. The produced last ideal value determines the succession best solution (universally) for current issue.
15. This process is rehashed for a population of Gene Sequences.
16. The optimized value of gene sequences is used for future analysis of disease (breast cancer).

### 4.3 Algorithm for MSADE

1. Initialize the population. Read mRNA shotgun sequences of both breast cancer disease and breast cancer suppressor gene sequences from NCBI site.
2. Start Multiple Sequence Alignment with required category of gene (DNA or RNA) input sequences.
3. Align pairwise sequence which belongs to same category of diseased or healthy sequences.
4. Create Distance matrix for each pair of sequences.
5. Optimal value will be generated based on alignment score.
6. Sequence generated relevant to optimal score will be considered as donor sequence and is used as input for DE.
7. Convert character formatted sequence to numeric format by substituting N, A, C, G, T values with 0, 1, 2, 3, 4.
8. Start mutation with three indiscriminate (random) vectors  $\times 1$ ,  $\times 2$ ,  $\times 3$  with indexed locations excluding current vector.
9. Find the difference  $x_{diff} = \times 3 - \times 2$  and create a new vector.
10. Multiply difference vector with mutation factor and add to  $\times 1$  vector then it generates current generation resultant vector.

$$[v = x_1 + F(x_2 - x_3)] \quad (4)$$

$v$ —donor vector,  $\times 1$ ,  $\times 2$ ,  $\times 3$  are three individuals from current generation,  
 $F$ —Mutation Factor, subtract individuals of current generation  $\times 2$  and  $\times 3$  and multiply with  $F$  and add to  $\times 1$  which results to form  $v$ .

11. Crossover is performed as part of recombination step on (resultant) donor vector which creates other third vector.
12. Apply the cost function on trial vector and target vector as part of Greedy selection step.
13. If trial vector score is fewer than target vector score then consider current trial vector as donor vector for next generation. If trial vector score is greater than target vector appends target vector score to generation vector score.
14. Find the generation average and generation best values and best solution vector for current generation for individual cycle.
15. Repeat steps 3–9 to get the optimal value for every one generation respectively.
16. Display the result and consider it for further categorization of gene sequences.

### 4.4 Flow Chart for MSAPSO

See Fig. 2.

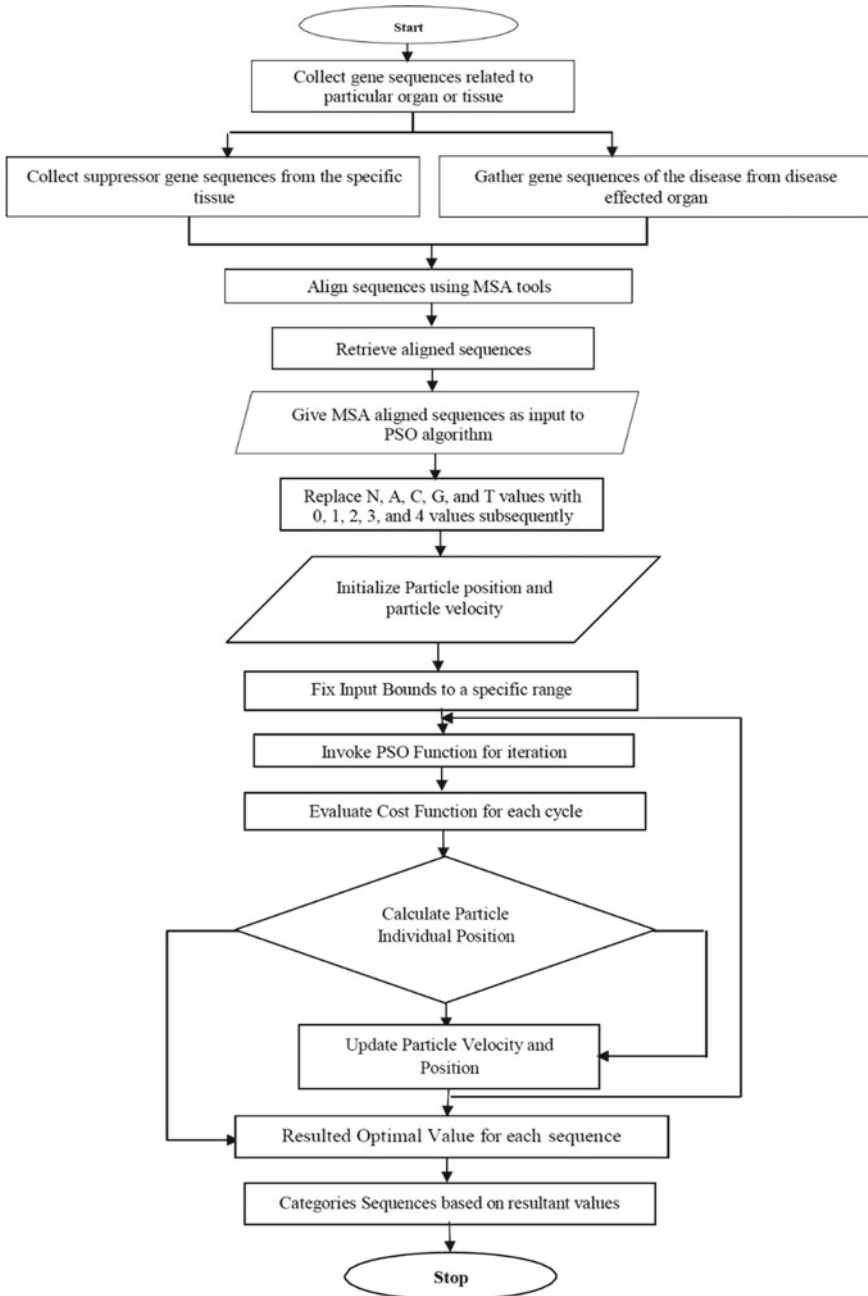


Fig. 2 Flow chart for multiple sequence alignment with particle swarm optimization

## 4.5 Flow Chart for MSADE

See Fig. 3 and Tables 1, 2.

## 5 Result Analysis

To perform any experiments on sequence databases reliable information sources are very important. Sequences are collected from different international sequence database sources. Those are Genbank at NCBI (USA), European Molecular Biology Laboratory (EMBL) at European Bioinformatics institute (EBI), and DNA Database of Japan (DDBJ) at National Institute of Genetics. These international databank sources maintain data conversant and maintain data under common protocol with same accession numbers in all databases. Computer and biological scientists can simply access public database as GenBank [6RA]. Data elements can be grouped based on diverse categories like as patient disease identifiers, patient selection based on disease criteria, treatment methods, and diagnosis of disease and conditions and end results [6RA2]. In the present work varied categories of sequence data base is collected based on patient disease such as cancer sequences, cancer suppressor gene sequences. The results are presented individually for diverse classifications of sequences after implementing hybrid algorithms i.e. MSA-PSO and MSA-DE. When the MSA-PSO implementation results are observed in classification of cancer aligned sequences these values are high compared to remaining cancer suppressor sequences. The similarity is also repeated on trial sequences of MSA-DE hybrid approach also. Compared to MSAPSO the results generated after MSADE is exhibiting more distinction when implemented on different classifications of sequences i.e. breast cancer sequences and breast cancer suppressor sequences.

### 5.1 Representation of Resultant Values in a Table Using Hybridized Algorithm MSAPSO

Figure 4 represents Optimal results of 10 sample MSA sequences after implementing MSAPSO on cancer gene sequences And graphical representation of related values. MSA aligned cancer sequences and optimal results generated after implementing MSAPSO hybrid algorithm on aligned sample sequences and related line chart is presented. The observed results are ranges in between maximum 1400 to minimum 900 approximately in case of cancer aligned sequences.

Figure 5 represents optimal results of 10 sample MSA sequences after implementing MSAPSO on cancer and cancer suppressor gene sequences

MSAPSO result of 10 sample aligned sequences and related line chart is presented. The observed MSAPSO results are ranges in between max. 750 and min. 130

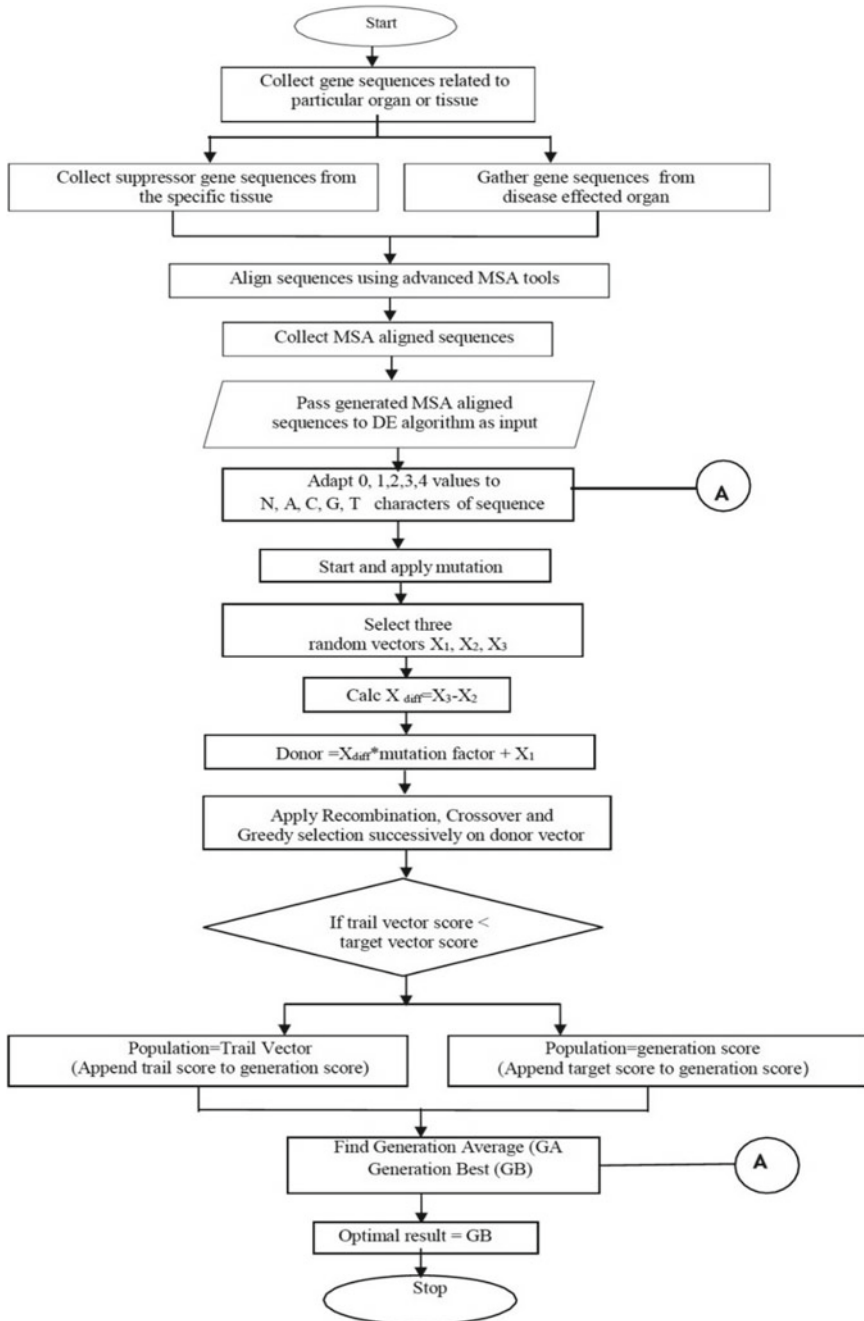


Fig. 3 Flow chart for multiple sequence alignment with differential evolution



**Table 1** Breast cancer sequences

S.No.	Accession No's
SEQ 1	KP255416.1 - TWH93401
SEQ 2	KP255415.1 - FP6401
SEQ 3	KP255414.1 - TWH37401
SEQ 4	KP255413.1 - TWH87101
SEQ 5	KP255412.1 - QMH17501
SEQ 6	KP255411.1 - TWH2001
SEQ 7	KP255410.1 - FP9401
SEQ 8	KP255409.1 - TWH69301
SEQ 9	KP255408.1 - HKSH5701
SEQ 10	KP255407.1 - FP3701
SEQ 11	KP255406.1 - HKSH12501
SEQ 12	KP255405.1 - FP4101
SEQ 13	AY150865.1
SEQ 14	AY093491.1 - IRCHF4B
SEQ 15	AF507077.1 - IRCHS6A
SEQ 16	AF507078.1 - IRCHS6B
SEQ 17	AY093486.1 - IRCHS7A
SEQ 18	AY093487.1 - IRCHS7B
SEQ 19	AY093488.1 - IRCHS16A
SEQ 20	AY144588.1

approximately in example of cancer suppressor aligned sequences. Figure 6 represents optimal results of 10 sample MSA sequences after implementing MSAPSO on cancer suppressor gene sequences

MSA aligned cancer and non-cancer sequences and optimal results generated after implementing MSAPSO hybrid algorithm on aligned sample sequences and related line chart is presented. The observed results are ranges in between maximum 600 and minimum 85 approximately in case of cancer and cancer suppressor aligned sequences.

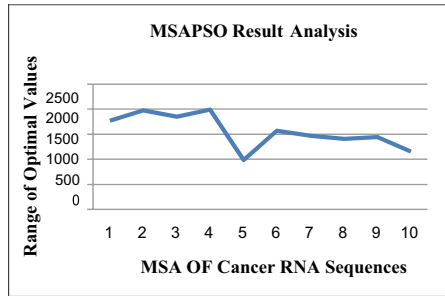
### 5.2 Representation of Resultant Values in a Table Using Hybridized Algorithm MSADE

Figure 7 represents MSA aligned cancer gene sequences. Implementation results of MSADE on 10 samples aligned sequences and related line chart is presented. The observed MSADE results are ranges in between maximum 14.00 to minimum 9.84 in case of cancer aligned sequences. Different categories of cancer sequence numbers are represented on X-axis and range of optimal values are represented on Y-axis.

**Table 2** Breast cancer suppressor

S.No.	Accession No
SEQ 1	AF209138.1
SEQ 2	AF066082.1
SEQ 3	AB700556.1
SEQ 4	AF209128
SEQ 5	AF209130.1
SEQ 6	AF209131.1
SEQ 7	AF209133.1
SEQ 8	AF209134.1
SEQ 9	AF209135.1
SEQ 10	AF209136.1
SEQ 11	AB699689.2
SEQ 12	AF209139.1
SEQ 13	AB699004.1
SEQ 14	AF209143.1
SEQ 15	AB118156.1
SEQ 16	AF209132.1
SEQ 17	AF209137.1
SEQ 18	AF209140.1
SEQ 19	EF178470.1
SEQ 20	AF209141.1

S NO	MSA ALIGNED CANCER SEQUENCES	MSAPSO RESULT
1	KP255416.1 & AF507077.1	1368
2	KP255415.1&KP255412.1	985
3	KP255414.1&KP255408.1	1766
4	KP255411.1&KP255407.1	1976
5	KP255410.1&KP255409.1	1847
6	KP255407.1&KP255406.1	1992
7	AY093486.1& AY093488	1572
8	AF507078.1&AY093487.1	1469
9	AY093486.1&AY093487.1	1146
10	AY093488.1&AY144588.1	1103

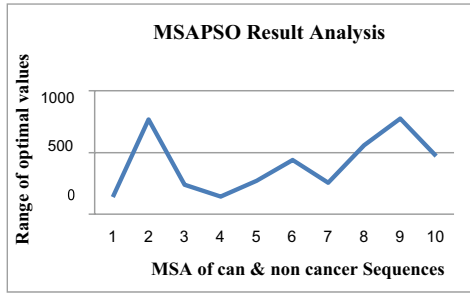


**Fig. 4** MSA aligned cancer sequences with its MSAPSO result and its related result analysis graph

Optimal results of 10 sample MSA sequences after implementing MSADE hybridized algorithm on cancer gene sequences are represented in Fig. 7. Figure 8 represents optimal results of 10 sample MSA sequences after implementing MSADE on cancer and cancer suppressor gene sequences.

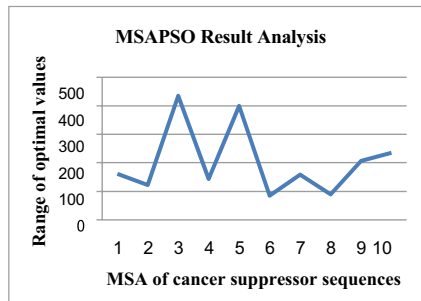
Implementation results of MSADE on 10 sample aligned sequences and related line chart is presented. The observed results after implementing MSADE are ranges in

S NO	MSA ALIGNED CANCER AND CANCER SUPPRESSOR SEQ'S	MSAPSO RESULT
1	KP255416.1&AF209135	138
2	KP255415.1&AF209135	766
3	KP255414.1&AF209138.1	236
4	KP255413.1&AF066082.1	142
5	KP255412.1&AB700556.1	260
6	KP255411.1&AF209128.1	439
7	KP255410.1&AF209130.1	154
8	KP255409.1&AF209131.1	560
9	KP255408.1&AF209133.1	774
10	KP255407.1&AF209134.1	550



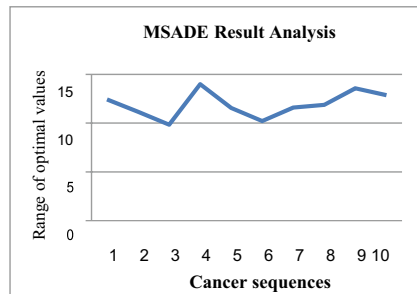
**Fig. 5** MSA aligned cancer diseased and suppressor sequences with its MSAPSO result analysis graph

S NO	MSA ALIGNED CANCER SUPPRESSOR SEQUENCES	MSAPSO RESULT
1	AF209138.1&AF066082.1	162
2	AB700556.1&AF209128.1	122
3	AF209130.1&AF209131.1	435
4	AF209136.1&AF209139.1	144
5	AB699689.2&AB699004.1	600
6	AF209140.1&EF178470.1	85
7	AF209132.1&AF209137.1	158
8	AB118156.1&AF209137.1	90
9	F209134.1 &AF209140.1	206
10	AF209130.1&AF209140.1	236



**Fig. 6** MSA aligned cancer suppressor gene sequences with its MSAPSO result analysis graph

DE implementation for CANCER sequences		
S No	DE MSA on cancer seq's	optimal value
1	AF507075.1,AY093484.1	12.42
2	AF507076.1,AF507077.1	11.17
3	AY093486.1,AY093487.1	9.84
4	AY093492.1,AY093493.1	13.99
5	KP255415.1,KP255412.1	11.54
6	KP255414.1,KP255408.1	10.21
7	KP255413.1,KP255410.1	11.61
8	KP255411.1,KP255407.1	11.85
9	KP255407.1,KP255406.1	13.56
10	AF507078.1,AY093487.1	12.86



**Fig. 7** MSA aligned cancer 137 sequences with its MSADE result analysis graph

between 4.95 and 0.33 in incident of cancer and cancer suppressor aligned sequences. Figure 9 denotes Optimal results of 10 sample MSA sequences after implementing MSADE on cancer suppressor gene sequences.

Implementation results of MSADE on 10 sample aligned cancer suppressor gene sequences and related bar graph is presented. The observed optimal results after

DE implementation for CAN-NONCAN sequences		
S no	MSA DE on CAN_NONCAN seq's	Optimal value
1	KP255416.1,AB699689.2	0.33
2	KP255415.1,AB699689.2	0.44
3	KP255414.1,AF066082.1	3.16
4	KP255413.1,EF178470.1	1.13
5	KP255412.1,AF209128.1	2.55
6	KP255411.1,AF209137.1	1.99
7	KP255410.1,AF209139.1	1.01
8	KP255409.1,AF209140.1	4.95
9	KP255408.1,AB118156.1	4.57
10	KP255407.1,AB700556.1	4.16

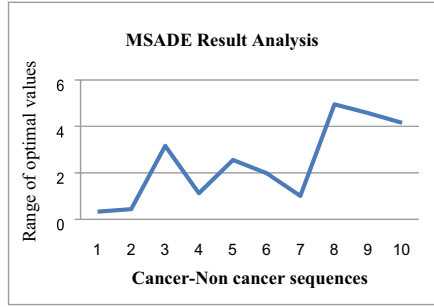


Fig. 8 MSA aligned cancer-non cancer sequences with its MSade result analysis graph

DE implementation for cancer suppressor gene sequences		
S no	DE_MSA on cancer suppressor seq's	optimal value
1	AF209138.1,AF066082.1	0.0001
2	AB700556.1,AF209128.1	0.0004
3	AF209140.1,AB118156.1	3.09
4	AF209130.1,AF209140.1	0.96
5	AF209134.1,AF209140.1	0.23
6	AF209136.1,AF209139.1	0.06
7	AB699689.2,AB699004.1	2.09
8	AB118156.1,AF209137.1	0.002
9	AF209132.1,AF209137.1	0.22
10	AF209140.1,EF178470.1	0.005

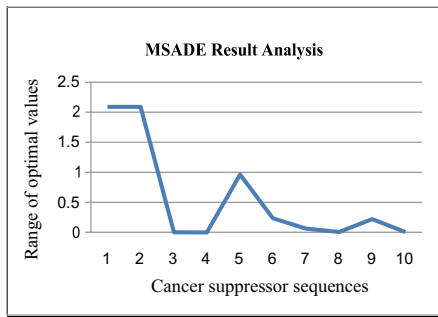


Fig. 9 MSA cancer suppressor gene sequences with its MSade result analysis graph

implementing MSade hybrid algorithm are lies in between maximum 4.95 to minimum 0.33 in case of cancer suppressor aligned sequences.

## 6 Conclusion

A hybrid approach is proposed using multiple sequence alignment with optimal search algorithms of soft computing techniques particle swarm optimization and Differential Evaluation. This approach is providing better results by providing more accurate optimal results as compared to simple PSO and DE implementation presented in previous work. In the proposed above observed values it is stated that compared to MSA-PSO implementation MSA-DE is providing more accurate optimal values which provides better categorization of diseased (breast cancer), healthy or suppressed gene sequences which aids in further investigates in future this hybrid methodology may be implemented on different categories of genetic diseases.

## References

1. J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, Cancer incidence and mortality worldwide international agency for research on cancer. IARC CancerBase No. 11. Lyon; France (2013). Accessed 1 Jan 2016. GLOBOCAN 2012 v1.0, <http://globocan.iarc.fr>
2. World Cancer Report, International Agency for Research on Cancer (2008). Retrieved 26 Feb 2011
3. Top 100 Cited Classic Articles in Breast Cancer Research, Eur. J. Breast Health **13**(3), 129–137 (2017). Published online 2017 Jul 1. <https://doi.org/10.5152/ejbh.2017.3480>
4. O. Hoffman, L. Fejerman, D. Hu, S. Huntsman, M. Li, E.M. John, G. Torres-Mejia, E. ZivEmail, Identification of novel common breast cancer risk variants at the 6q25 locus among Latinas. ORCID, profile Breast Cancer Research 201921:3 <https://doi.org/10.1186/s13058-018-1085-9> © 4 December 2018. Published: 14 Jan 2019
5. P. Apostolou, F. Fostira, Hereditary breast cancer: the era of new susceptibility genes. Biomed. Res. Int. 747318 (2013). Published online 2013 Mar 21. <https://doi.org/10.1155/2013/747318>. PMID: PMC3618918 PMID: 23586058
6. T. Walsh, M.K. Lee, S. Casadei, A.M. Thornton, S.M. Stray, C. Pennil, A.S. Nord, J.B. Mandell, E.M. Swisher, M.-C. Kinga, Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc. Nat. Acad. Sci. USA **107**(28), 12629–12633 (2010). Published online 2010 <https://doi.org/10.1073/pnas.1007983107>. PMID: PMC2906584 PMID: 20616022
7. S. Shiovitz, L.A. Korde, Genetics of breast cancer: a topic in evolution. Ann. Oncol. **26**(7), 1291–1299. Published online 2015 Jan 20. <https://doi.org/10.1093/annonc/mdv022>. PMID: PMC4478970 PMID: 25605744
8. S. Mitra, Y. Hayashi, Bioinformatics with soft computing. **36**(5), 616–635 (2006). Date of Publication: 21 Aug 2006 ISSN Information: INSPEC Accession Number: 9049333 <https://doi.org/10.1109/tsmcc.2006.879384>
9. R.K. Jena, M.M. Aqel, P. Srivastava, P.K. Mahanti, Soft computing methodologies in bioinformatics. Eur. J. Sci. Res. **26**(2), 189–203 (2009). <http://www.eurojournals.com/ejsr.html>, ISSN 1450-216X
10. L.N.J. Gavarraju, P. Jeevana Jyothi, K. Karteeka Pawan, A literature survey on multiple sequence alignment algorithms. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **6**(3) (2016). ISSN: 2277 128X
11. K. Lohitha Lakshmi, P. Bhargavi, S. Jyothi, Soft computing techniques for gene annotation. Int. J. Latest Eng. Manag. Res. (IJLEMR) **03**(02), 26–34 (2018). [www.ijlemr.com](http://www.ijlemr.com), ISSN: 2455–4847
12. K. Bhargavi, S. Jyothi, Classification of DNA sequence using soft computing techniques: a survey. Indian J. Sci. Technol. **9**(47). <https://doi.org/10.17485/ijst/2016/v9i47/89343>, ISSN (Print): 0974-6846, ISSN (Online): 0974-5645
13. K. Lohitha Lakshmi, P. Bhargavi, S. Jyothi, An analysis of breast cancer DNA sequences using particle swarm optimization. Int. J. Eng. Technol. **7**(4.7), 335–338 (2018). [www.sciencepubco.com/index.php/IJET/Researchpaper](http://www.sciencepubco.com/index.php/IJET/Researchpaper)
14. S. Cheng, Y. Shi, Q. Qin, Population diversity based study on search information propagation in particle swarm optimization, in *Proceedings of 2012 IEEE Congress on Evolutionary Computation, (CEC 2012)* (2012). IEEE, Brisbane, Australia, pp. 1272–1279
15. A. Kamal, A. TarekSayed, M. Mahroos, *Sequence Alignment Using Parallel Particle Swarm Optimization*. Content uploaded by Mohsen Mahroos (2012)
16. Z. Ibrahim, L.K. Sheng, F. Naim, M.F.M. Jusof, N.W. Arshad, An analysis of archive update for vector evaluated particle swarm optimization. Int. J. Intell. Syst. Appl. Eng. Adv. Technol. Sci. [www.atscience.org/IJISAE](http://www.atscience.org/IJISAE), Accepted 15th Aug 2014, <https://doi.org/10.18201/ijisae.48588>, ISSN: 2147-67992147-6799
17. D.R. Penas, J.R. Banga, P. Gonz'alez, R. Doallo, A parallel differential evolution algorithm for parameter estimation in dynamic models of biological systems, ed. by J. S'aez-Rodr'iguez

- et al. *8th International Conference on Practical Application of Computer 173 Biology and Bioinformatics (PACBB 2014), Advances in Intelligent Systems and Computing 294*. [https://doi.org/10.1007/978-3-319-07581-5\\_21](https://doi.org/10.1007/978-3-319-07581-5_21), © Springer International Publishing Switzerland 2014
18. K. Lohitha Lakshmi, P. Bhargavi, S. Jyothi, *An Analysis of Breast Cancer Gene Sequences using Differential Evaluation*. <https://doi.org/10.5013/ijssst.a.20.01.35.35.1>, ISSN: 1473-804x online, 1473-8031 print
  19. M. Chatzou, C. Magis, J.M. Chang, C. Kemena, G. Bussott, Multiple sequence alignment modelling: methods and applications. *Brief. Bioinf.* (6) (2015). <https://doi.org/10.1093/bib/bbv099>
  20. F. Sviatopolk-MirskyPais, P. de CássiaRuy, G. Oliveira, R.S. Coimbra, Assessing the efficiency of multiple sequence alignment programs. *Algorithms Mol. Biol.* **9**, 4 (2014). Published online 2014 Mar 6. <https://doi.org/10.1186/1748-7188-9-4>
  21. T.O. Ting, X.-S. Yang, S. Cheng, K. Huang, *Hybrid Meta Heuristics Algorithms: Past, Present, and Future*. First Online: 28 Dec 2014
  22. S. Lalwani, R. Kumar, N. Gupta, A review on particle swarm optimization variants and their applications to multiple sequence alignment. *J. Appl. Math. Bioinf.* **3**(2), 87–124 (2013). ISSN: 1792-6602 (print), 1792-6939 (online)

# Performing Image Compression and Decompression Using Matrix Substitution Technique



T. Naga Lakshmi and S. Jyothi

**Abstract** Now a days, large amount of information storage and transmission is common in public sectors as well as private sectors like Governments, military, and so on. With increase of demand in digital network more and more accurate images are transmitted. Whenever we are transmitting an image with larger size it occupies more space. In order to reduce the size, Image compression an efficient and advantageous technique is used to remove the redundant data from the image. Thereby the storage space in memory is reduced and also it takes less time to transmit the image. Therefore, development of efficient image compression techniques has become necessary. In this paper, a new algorithm matrix substitution technique is proposed in order to reduce the occupied space by replacing group of bits with a single bit.

**Keywords** Image compression · Image decompression · Matrix substitution · Image storage · Image transmission

## 1 Introduction

Now a days, the transmission of digital Images has become more important in this present environment. The data storage and transmission plays a critical role. Hence the image with less size must be transmitted over the network, occupies less storage. The main functionality of image compression is to reduce the size by rearranging the input image pixels to preferred compression level. The image consists of pixel values where the adjacent pixels have same values that leads to spatial redundancy. If we group the redundant pixels with a single value the storage space is reduced. To

---

T. Naga Lakshmi (✉)

Department of Computer Science and Engineering, AITS, Rajampet, India

e-mail: [lakshmi.nag04@gmail.com](mailto:lakshmi.nag04@gmail.com)

S. Jyothi

Department of Computer Science, SPMVV, Tirupati, India

e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,

Learning and Analytics in Intelligent Systems 15,

[https://doi.org/10.1007/978-3-030-46939-9\\_3](https://doi.org/10.1007/978-3-030-46939-9_3)

achieve this, we require a compression technique with little amount of loss to achieve high Compression ratio that maintains good quality of decompressed image.

Variety of lossy and lossless compression algorithms adaptable for multidimensional data compression are discussed by Jain [1]. Lossy compression techniques provide high compression rate, but it is difficult to get an exact data. Whereas lossless methods recover the original data exactly, but it cannot compress the image to maximum level as defined by the former method. So, when there is loss in the data which can be measured as negligible, the compression ratios can be maximized. When a compressed image is decompressed to the original one, it occurs some loss, but it is acceptable without compromising. We can achieve good quality of image because Human Visual System does not detect the little bit changes in the image discussed by Marta Mark and Grgic [2]. The method that has been used by Alexander et al. [3] gives good results with a limitation of hardware dependency. The image is segmented into blocks with various sizes for compression by using neural networks by Vilovic [4]. The results show the usage of multilayer perceptron's for image compression. The algorithms for restoration of images with poor dynamic range is discussed by Caselles et al. [5]. Various approaches exist in literature which propose interpolation techniques for "perceptually motivated" coding applications [6–8].

The purpose of this paper is to propose a new algorithm for performing the compression of digital image and then decompressing the compressed one with little loss which is negligible, compared to the original image. We use a sub matrix substitution technique replaces the entire sub matrix with the single bit. The same process is continued for entire sub matrices. The proposed decompression algorithm can also be universally used to enlarge any given image with the single bit of sub matrix values with some loss of pixels.

## 2 Compression of an Image

Image Compression is a technique used to reduce the size of an image without degrading the quality of an image to a maximum extent. To reduce the size an image is encoded with few bits by eliminating the redundant bits without affecting the quality of an image. Two kinds of image compression techniques are there: lossy and lossless compression methods. In lossy compression technique the part of an image is removed which occurs loss to the original image such as JPG, GIF etc., In lossless compression the redundant part is removed without affecting the quality of an image such as BMP, PNG, RAW etc.

Consider a gray scale image of size  $M \times N$  to perform the compression, 0 represents to black and 255 represents to white. The values between 0 and 255 are the variants of black and white. The original matrix can be shown as in Fig. 1. The two sub matrices are chosen and one pixel is selected forms a new image with two selected pixels shown in Fig. 2.



	1	2	3	4	5	6	7	8	9
1	176	174	174	176	180	185	193	184	184
2	177	177	177	178	180	182	185	184	183
3	186	182	179	177	176	175	176	171	175
4	176	177	179	181	183	185	188	180	173
5	172	173	174	175	176	178	180	174	169
6	172	171	172	174	178	184	192	176	176
7	167	168	170	171	172	174	175	168	170
8	182	171	164	160	160	163	170	176	179
9	173	171	169	168	167	167	167	169	172

Fig. 1 Image in matrix form

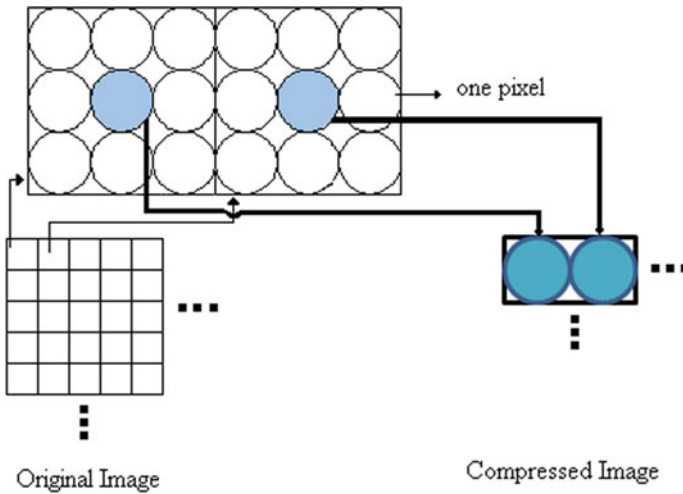


Fig. 2 Image submatrix representation

**Algorithm for matrix substitution compression**

1. Check whether the size of the matrix  $M \times N$  is multiple of 3 or not. If not resize the image of size  $X \times Y$  by preprocessing the original image. This can be done by padding the bits in two ways
  - (i) Pad the empty pixel positions with zero's with little bit error rate
  - (ii) Replace the empty pixels with the surrounded pixel values with less error rate which is negligible
2. Divide the image (matrix) into sub matrices of size  $K \times K$  where  $K$  is the multiple of 3
3. Replace the submatrix elements with a single pixel value as shown in Fig. 2

- Repeat the same process for all sub matrices, replaced with their respective single pixel values.

Suppose consider the matrix of size  $9 \times 9$  and it is divided into 9 submatrices which can be shown in Fig. 3. The shaded are the first 3 sub matrices, only one pixel is highlighted which can be used to replace the submatrix values.

The same procedure is applied for all the sub matrices of an original image. In each matrix the middle element has chosen, and it is replaced with the single pixel value as shown in Fig. 4. There by a new matrix of size  $X \times Y$  is formed after performing the compression.

	1	2	3	4	5	6	7	8	9
1	176	178	178	177	176	175	184	192	191
2	181	176	172	173	176	176	185	193	192
3	186	172	166	170	178	177	185	193	191
4	184	173	168	172	179	178	185	191	188
5	178	177	177	177	178	178	182	185	184
6	180	183	185	181	175	172	175	180	182
7	190	188	185	180	174	165	167	176	180
8	188	186	183	180	177	173	174	176	175
9	180	180	182	182	181	181	182	181	175

Fig. 3 Image in matrix form pixels are selected

	1	2	3	4	5	6	7	8	9	10
1	176	176	193	184	183	178	183	178	184	199
2	177	178	185	184	181	180	167	178	183	179
3	186	177	176	171	181	181	173	173	187	182
4	176	181	188	180	167	177	182	178	184	177
5	172	175	180	174	168	187	182	180	179	178
6	172	174	192	176	175	175	184	176	180	183
7	167	171	175	168	175	185	164	191	176	178
8	182	160	170	176	181	174	178	170	169	176
9	173	168	167	169	174	165	171	169	170	177
10	188	168	179	178	176	175	184	178	181	173

Fig. 4 Compressed matrix values

### 3 Decompression of an Image

Decompression is the process of restoring the compressed image into its original form. It is widely used for transmitting the images in military, governments etc., it may be lossy or lossless depends on the technique that we are using.

The decompression method is applied for the compressed image of size  $X \times Y$ , all the pixel values are formed from the submatrices. To decompress the image, we need to replace the single bit values with sub matrix elements. The pixel values are interpolated with new values. Each and every pixel has two values, one is pixel at position  $(a, b)$  and the other is the pixel value i.e., gray scale value  $G$ . consider  $r$ th pixel at position  $(a_x, b_y)$  on  $\beta$ -axis and has pixel value (i.e., gray value) as  $G_r$  on  $G$ -axis which can be shown in Fig. 5.

**Algorithm for decompression**

Consider the compressed image  $X \times Y$  shown in Fig. 4, let's start the implementation by considering the first pixel at position  $(1, 1)$ .

1. Get the pixel values  $P_1, P_2, P_3$  at the positions  $(1, 1), (1, 2), (1, 3)$  respectively by moving in  $X$ -axis.
2. Here  $P_1, P_2, P_3$  are the pixels need to be enlarged with their respective sub matrices.
3. Obtain gray values for the positions  $P_1, P_2$  and  $P_3$  are  $G_1, G_2$ , and  $G_3$  respectively.
4. From the Fig. 2 two pixels are there between two consecutive compressed pixels. We need to interpolate these two pixels it can be done as follows:
  - i. The positions of  $P_1, P_2$  and  $P_3$  were plotted on graph with  $G$  and  $\beta$  axis shown in Fig. 6.  $P_1$  is at position  $(1, G_1)$  and  $P_2$  is at  $(1 + 3, G_2)$  and  $P_3$  is at  $(1 + 6, G_3)$  and so on.
  - ii. The equation for calculating the interpolated pixels is

$$18G = (G_1 - 2G_2 + G_3)\beta^2 + (11G_1 + 16G_2 - 5G_3)\beta + 28G_1 - 14G_2 + 4G_3 \tag{1}$$

**Fig. 5** Representation of pixel on  $(\beta, G)$  axis

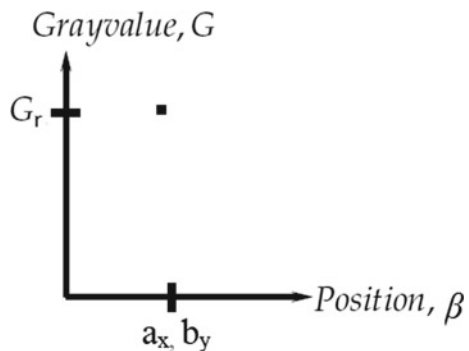
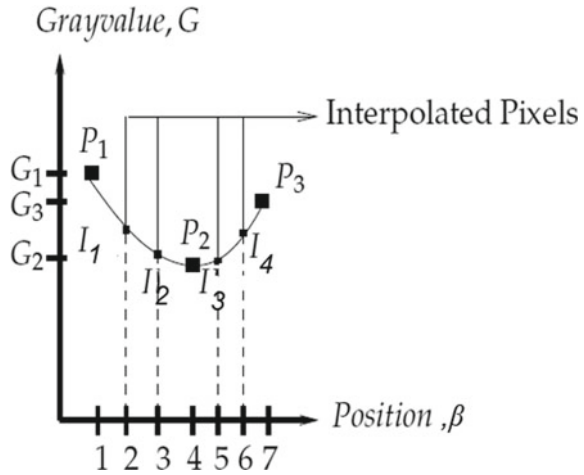


Fig. 6 Derived curve



- iii. By using Eq. (1), interpolated pixels I1, I2, I3 and I4 are calculated with respect to 2, 3, 5 and 6 positions on  $\beta$ -axis respectively.
  - iv. Obtain Ig2, Ig3, Ig5 and Ig6 values from I1, I2, I3 and I4 as shown in Fig. 6
  - v. Rearrange the pixels as G1 Ig2 Ig3 G2 Ig4 Ig5 G3 at positions (1, 1) (1, 2) (1, 3) (1, 4) (1, 5) (1, 6) and (1, 7) produces a new matrix
5. Consider the next two adjacent compressed pixels are replace with the interpolated pixels. The same procedure is considered in x-axis.
  6. After completion of horizontal interpolation, the same procedure is repeated for y-axis (vertically) produces a final matrix with original size.

After performing the same operation vertically, the original image has been formed and it can be shown as matrix form in Fig. 7.

## 4 Computational Results

The overall procedure of compression to decompression can be shown in Fig. 8. Firstly the compressed image is distributed horizontally and again distributed vertically. All the interpolated pixels are arranged in the matrix produces an original matrix.

### 4.1 Different Compression Ratios of an Image

Consider an image shown in Fig. 9 the image is cropped as shown in Fig. 10a and different decompressed ratios are applied for the cropped part. Figure 10b–d depicts

	1	2	3	4	5	6	7	8	9
1	176	174	174	176	180	185	193	184	184
2	177	177	177	178	180	182	185	184	183
3	186	182	179	177	176	175	176	171	175
4	176	177	179	181	183	185	188	180	173
5	172	173	174	175	176	178	180	174	169
6	172	171	172	174	178	184	192	176	176
7	167	168	170	171	172	174	175	168	170
8	182	171	164	160	160	163	170	176	179
9	173	171	169	168	167	167	167	169	172

Fig. 7 Decompressed matrix

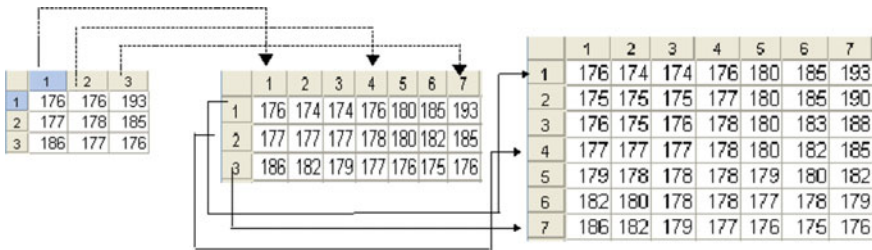
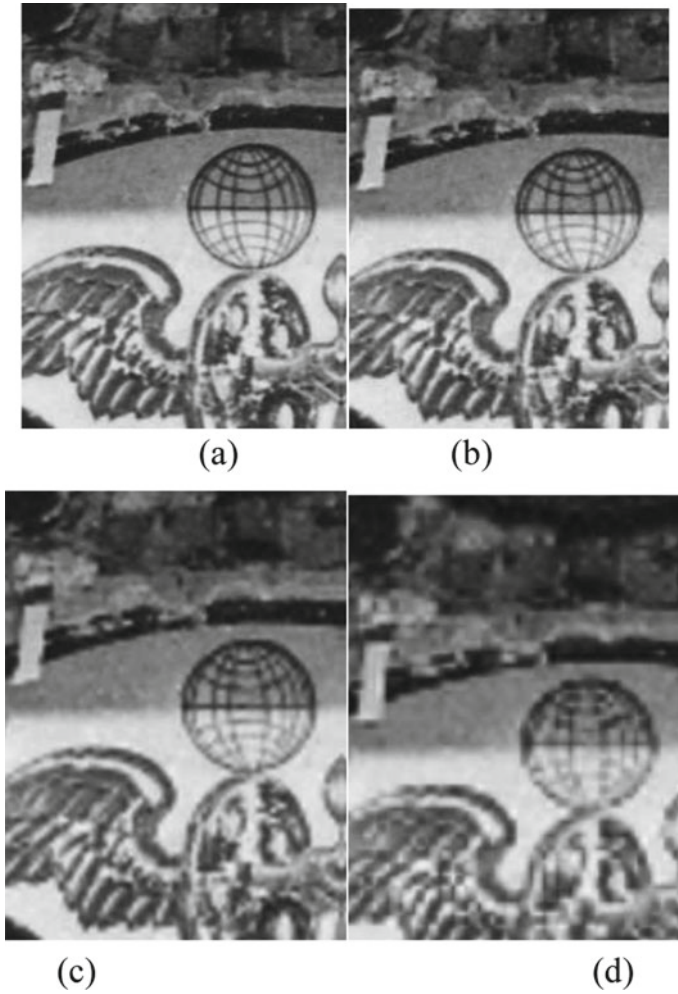


Fig. 8 The process of compressed image to decompressed image



Fig. 9 Original image



**Fig. 10** Comparison of image with different compression ratios

the decompression with 4:1, 8:1 and 16:1 respectively. If we are applying 32:1 compression ratio, our decompression algorithm enlarges the image and produces blurred image. The compression ratio is calculated from uncompressed size to compressed size and the equation is

$$\text{Compressed Ratio} = (\text{uncompressed size}/\text{compressed size}) : 1$$

The recommended compression ratios for different size of images can be shown in Table 1. It can be increased when the size of the file increases. Some of ranges

**Table 1** Various image sizes with recommended compression rates

Image size	Recommended compression ratio
1–850 KB	4:1
850 KB–6 MB	8:1
6–10 MB	16:1

can be shown in Table 1 the images will blur if we go for more compression rate for a given range.

### 4.2 Output Results

The Original image is given as input to the proposed algorithm can be shown as Fig. 11 produces the decompressed image can be shown in Fig. 12. All the images considered are raw images only.

The decompressed image is like the original images with minimum loss of data which is negligible even human vision not able to detect the variation.

The comparison of different images is shown in Table 2. Different images with vary sizes are given as inputs to our proposed algorithm. The sizes are compared and

**Fig. 11** Original image



**Fig. 12** After decompression





**Table 2** Comparison of different image sizes

Image size	Compressed size (KB)
32.2 MB	568.5
27 MB	474.5
469.4 KB	101.1

**Table 3** Error estimation

Image name	Original size	Compressed size (KB)	Error rate (mean square error)
Image 1	32.2 MB	568.5	$5.0122e^{-005}$
Image 2	27 MB	474.5	$1.3241e^{-005}$
Image 3	469.4 KB	101.1	$2.8355e^{-004}$

examined for various input sizes, the compressed size for the inputs are as shown in Table 2.

### 4.3 Error Measurement

The most important method used for assessing the quality of an image is measuring the error. It provides high correlation to image quality. Common method used for error measurement is “Mean Square Error”. The average mean square error for an  $P \times Q$  image is defines as:

$$e_{ms}^2 = \frac{1}{PQ} \sum_{i=1}^P \sum_{j=1}^Q (X_{i,j} - X'_{i,j})^2 \quad (2)$$

where  $X_{i,j}$  and  $X'_{i,j}$  represent the  $P \times Q$  original and the reproduced images, respectively. The average mean square error is often estimated by average sample mean-square error. Our proposed method is lossy reproduces the image with degradation, but it is not noticeable for the human eye. We can also use some other techniques for reducing the error rate. For different images the compressed size and the error estimation is shown in Table 3.

### 4.4 Comparison of Results

Our proposed method is compared with other compression techniques. Table 4 illustrates our proposed technique is compared with respect to JPEG. As observed from the test results, when the input image is large in size our algorithm provides good



**Table 4** Comparison with JPEG method

Image name	Original size	JPEG size	Our compression method (KB)
Image 1	32.2 MB	2.1 MB	568.5
Image 2	27 MB	1.1 MB	474.5
Image 3	469.4 KB	98.3 KB	101.1

compression. But for the smaller images our algorithm compresses less than JPEG. This algorithm is best suitable for images with larger size.

Some of the different images are collected and applied our substitution compression and decompression algorithm produces the outputs as shown in Fig. 13. As we observed from the test results the input size also plays a role for the compression. Our decompression algorithm can also be applied for image enlargement. If we are enlarging for more size the input image will gets blurred. But this is not applicable for zoom in operation again a new algorithm has to be implemented. We can also use more number of images for testing the proposed algorithm.



**Fig. 13** Original image versus decompressed images

## 5 Conclusion

This paper proposed a new method called submatrix compression algorithm, computationally less expensive to reduce the image size. The decompression algorithm depends on interpolation of gray scale pixels, applied to the decompressed image. Thus, provides a good result to compress an image. It also compares the compression rates for different sizes of images. As noticed from the restored image it is almost same as original image and the loss is negligible. The comparisons with different images have shown in Table 2. The decompression algorithm enlarges the compressed image to original one by substituting the pixels from interpolation. The techniques suggested here are suitable for applications that can be easily transmitted over the network. Our decompression algorithm used for any kind of decompressed image with minimum loss. This algorithm is suitable for large sized images produces good compression. The decompression algorithm can also be used for zoom out operation. The limitation in our proposed algorithm is if the image size is small it doesn't provides good compression.

## References

1. A.K. Jain, Image data compression: a review. *Proc. IEEE* **69**(3), 349–389 (1981)
2. S.G. Marta Mark, M. Grgic, Picture quality measures in image compression systems, in *Eurocon* (2003), pp. 233–236
3. L.T.W. Alexander, P. Morgan, R.A. Young, A gaussian derivative based version of jpeg for image compression and decompression. *IEEE Trans. Image Process.* **7**(9), 1311–1320 (1998)
4. I. Vilovic, An experience in image compression using neural networks, in *48th International Symposium ELMAR-2006, Zadar, Croatia, 07-09 (2006)*, pp. 95–98
5. V. Caselles, An axiomatic approach to image interpolation. *IEEE Trans. Image Process.* **7**(3), 376–386 (1998)
6. J.R. Cass, Image compression based on perceptual coding techniques, Ph.D. dissertation, Dept. Signal Theory Commun., UPC, Barcelona, Spain (1996)
7. H.L. Floch, C. Labit, Irregular image subsampling and reconstruction by adaptive sampling, in *Proceeding of International Conference of Image Processing ICIP*, vol. III, Lausanne, Switzerland (1996), pp. 379–382
8. X. Ran, N. Favardin, A perceptually motivated three-component image model. part ii: Applications to image compression. *IEEE Trans. Image Process.* **4**, 430–447 (1995)

# Classification of Cotton Crop Pests Using Big Data Analytics



R. P. L. Durgabai, P. Bhargavi, and S. Jyothi

**Abstract** Agriculture is the main occupation in Andhra Pradesh (A.P.), the atmosphere and land of A.P. is appropriate for cultivating variety of crops like rice, wheat and cotton. Crop protection is one of the foremost challenges in agriculture. Pest classification during unusual weather conditions is very strong confront and the goal of every farmer is to protect the crops in exact time. Big Data analytics is playing a dominant role in agriculture sectors which helps in making good decisions to prevent any crops loss. Machine learning algorithm is a best analytical platform that automates analytical model building. Classification is a main method of machine learning which is useful to classify the problem and provide better solutions. This paper, demonstrates the implementation of three classifiers K-Nearest Neighbor (K-NN), Naive Bayes (NB), Decision Tree (DT) on cotton pests data, out of which the decision tree classifier proved to be the best for analysis.

**Keywords** Cotton · Pests · Classification · Big data · Decision tree · Machine learning

## 1 Introduction

There will be a crop loss due to various reasons, but pests are one of the important factors which influence the growth of crop in all stages. Loss of cotton yield occurs due to pests and it affects the world annual production up to 15%. The major pests that attacked the cotton crop from 1992 to 2017 in A.P. are Aphids, Jassids, Thrips, Whitefly, Leafhopper and Pink Boll worm. Analysis of pests of cotton in different weather conditions is dealt using big data analytics [1].

---

R. P. L. Durgabai (✉) · P. Bhargavi · S. Jyothi  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati 517502, India  
e-mail: [poonamramchandra@gmail.com](mailto:poonamramchandra@gmail.com)

P. Bhargavi  
e-mail: [pbhargavi18@yahoo.co.in](mailto:pbhargavi18@yahoo.co.in)

S. Jyothi  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_4](https://doi.org/10.1007/978-3-030-46939-9_4)

Big Data is a phrase used to define a large amount of structured and unstructured data which is difficult to process and analyze in traditional database. Analytics of such a massive data will be useful for better decision making to farmers. To get the best result analysis of big data, machine learning is required. The existing data set is gigantic and for better conclusion decision tree classifier has been implemented [2].

Agriculture is the strength of the Indian economic system. It stands as the main pillar and livelihood in India. Now-a-days agriculture is made easy because of advancements and latest technologies like sensors, devices, machines and information technology. Agriculture includes cultivation of many crops, out of which the major share of farming goes to cotton in A.P. Cotton is the second largest crop in India. Andhra Pradesh is playing an imperative role in national efforts for rising up of cotton production and productivity in India. It is a khariff crop and it is sown in August–September, it is affected by different pests in various climatic conditions [3].

This paper presents, classification of cotton pests based on weather parameters using decision tree classifier, which is a most suitable machine learning model for classification.

## 2 Literature Survey

Analysis by researchers proved that pests classification help farmers to save the crop in right time. According to Revathi et al. [4] bring out the image RGB ranging techniques used to identify the diseases in which, the captured images are processed for enhancement first. Patil et al. [5] presented an intelligent system for effectual prediction of pest population of dynamics of thrips on cotton crop. Sagar et al. [6] reviewed about role of mixed functions oxidizes that plays a major role in development of insecticide resistance in cotton leaf hopper. Dey et al. [7] demonstrates an automated approach for detection of white fly pest from leaf images of various plants. Rajan et al. [8] describes the study of various image processing techniques and applications for pest identification and plant disease detection. Das et al. [9] identified the visual symptoms of plant diseases by means of a machine vision system and providing a solution for disease control. Javed et al. [10] interprets the image which has been segmented from the fields by using enhanced K-Mean segmentation technique that identifies the pest or any object from the image. Osisanwo et al. [11] expounds the classification and comparison among various supervised machine learning algorithms. Badage [12] illustrates the early detection of diseases as soon as it starts spreading on the outer layer of the leaves. Akila et al. [13] proposed a deep learning based approach to detect leaf diseases in many different plants using images of plant leaves.

### 3 Problem Domain

A Problem domain is giving the impression of being at only the area of person’s interest, and the theme of capability or application that needs to be study to work out a problem. The present problem is to work out the classification of various pest of cotton crop all through different weather conditions.

#### 3.1 Description of Cotton Pests Dataset

The Principal Scientist, Department of Entomology at the Agriculture University, Guntur, A.P., India, collected the observation data to understand the six major pests population on cotton crop throughout different weather conditions in Andhra Pradesh.

Table 1 explains parameters of the present data set, that holds year, set of weekly recordings regarding the weather and pest occurrence in the state and the data recording range from the year 1992–2017. A brief explanation of the different attributes in the data is as follows: Year—The year is mentioned to notice in which type of pests attacked in which year. Standard week—weeks in each year and their pests occurrence based on weather parameters. Weather conditions: Temperature high (°C): The highest temperature or the maximum temperature recorded (Temp. H). Temperature Low(°C): The lowest temperature or the minimum temperature recorded (Temp. L). Humidity: The relative humidity recorded in morning (RH I (%)) and evening (RH II (%)), Rainfall (mm): The amount of rainfall recorded in each week. Average Pest (Pestavg): The average pest gives the total pests occurrence in each year.

**Table 1** Features of cotton crop dataset

S. No.	Parameters
1	Year
2	Standard week
3	Aphids
4	Jassids
5	Thrips
6	Whitefly
7	Leafhopper
8	Pinkboll worm
9	Temp. H
10	Temp. L
11	RH I
12	RH II
13	Rainfall
14	Pestavg

## **4 Methodology**

Methodology is the specific practice or method which is used to analyze information about an area. The current data is analyzed by different classifiers of machine learning technique.

### **4.1 *Big Data Analytics***

In this technological world, voluminous data which is generated every fraction of second can be dealt through data analytics which is playing a key role in all organizations for better and profitable decision making. Big data is categorized into three important characteristics such as velocity, volume and variety. The traditional databases are not in use as they are unable to analyze and process variety of data. Agriculture is one of the most important sectors in India where majority of people depend on it for their livelihood. In some states agriculture is the main profession. The data analytics will help to analyze and can make better predictions of the pest occurrence that leads to enormous loss which not only effects the crop but also farmers and their dependents [14]. The data is stored in Mongo DB which is a document based NoSQL database system.

Mongo DB is a simple, dynamic and helpful to store enormous data, it will use JSON or BSON like documents with schema. It is to implement a data store that provides high performance, availability, scalability and flexibility. The data is stored as a separate document inside a collection, instead of storing in rows and columns of a traditional relational database [15]. The present dataset is converted into Java script object notation format and stored in mongo db with AP cotton database name and a collection name is cotton pest with more than one million numbers of samples.

### **4.2 *Classification of Cotton Crop Pests***

Classification belongs to supervised learning branch of machine learning, in which system program learns from the data input given to it and then uses this learning to classify new observation.

### **4.3 *Machine Learning***

Machine learning is an area of artificial intelligence that aims at facilitating machines to execute their tasks by using intellectual software. Machine learning is a trending technology and it can be used in modern agriculture industry to find solutions to the

problems. Machine learning can be divided into three categories, namely supervised learning, unsupervised learning, reinforcement learning. Supervised learning aims to analyze the training dataset and produces a correct outcome from labeled data. The three main classifiers like K-NN, Naive Bayes and Decision tree are applied on the current dataset.

### **K-Nearest Neighbor**

K-NN classifier belongs to the supervised machine learning family, which contains certain advantages such as it can implement multi-class problem, and it is pretty perceptive. When it is applied on the current data set it is proved that the performance time is too slow, and it is difficult to maintain homogeneous data throughout the dataset which is the characteristic feature of this algorithm. Finally it is observed that the accuracy obtained through K-NN algorithm is less compared to decision tree classifier.

### **Naive Bayes**

Naive Bayes classifier is a probabilistic machine learning algorithm which is used for classification task. It is simple to implement and best suited to any dataset. As the current dataset is massive the accuracy obtained is very low compared to other two classifiers and it is taking much time to fit the data.

### **Decision Tree**

The decision tree classifier is applied on the current dataset as it is able to classify the pest based on weather conditions, which makes a naive person to understand easily, it is useful in data examination, the classifier performed on all the required features and variable screening is also done, the test results has been generated and analysis is visualized in the form of tree structure, interpretation helps to take wise decisions. In Machine learning algorithms, the dataset used to provide for the model is generally divided into training, validation and test phases. The training process evaluates the model accuracy in both phases. Once the model has been trained and is ready to use, a data set shall be selected to test the model under different conditions. In this study, the models were trained by using the weather parameters like temperature high, low, relative humidity I, relative humidity II and rainfall from 1992 to 2017. This training dataset was randomly divided as 80% for training and 20% for test. Classification of the cotton crop pests based on the weather conditions is modeled and tested with samplings of twenty five years.

## **5 Experimental Results**

The classifiers are applied on cotton pest dataset to classify different pest occurrences based on weather conditions.

**Table 2** Comparative study of classification algorithms

S. No.	Classifier applied	Accuracy	Prediction time in seconds	Rules generated	Interpretation	Visualization
1	K-NN	58.61	776	No	No	No
2	Naive Bayes	37.27	372	No	No	No
3	Decision tree	65.81	300	Yes	Yes	Yes

Table 2 describes comparative study of classification algorithms each classifier and its accuracy and time taken to predict each model. Comparatively Decision tree can make clear analysis and it is able to generate rules which help for interpretation and visualization.

### 5.1 Decision Tree Classifier for Pest Classification

In Machine learning there are many classifiers but decision tree is chosen in the present dataset as the agriculture sector can make decisions to achieve the target. Decision tree constructs classification models in the shape of a tree formation. It splits down a dataset into smaller and smaller subsets while on the similar moment a connected decision tree is gradually acquire. The concluding outcome is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node signify a classification or decision. The best predictor is called the root node which is a topmost decision node in a tree. It can handle both categorical and numerical data [16].

### 5.2 Test Results and Analysis

Decision tree classified the current dataset by taking all the independent variables such as X0 = Maximum Temperature (Temp. H), X1 = Minimum Temperature (Temp. L), X2 = Relative Humidity in morning (RH I), X3 = Relative Humidity in evening (RH II), X4 = Rainfall, and dependent variables are types of pests with values in an order 0 = Aphids, 1 = Jassids, 2 = Thrips, 3 = Whitefly, 4 = Leafhopper, 5 = Pink boll worm. Gini is a criterion to reduce the feature impurity. It has generated the following rules to classify the pests (Fig. 1).

Rule 1 If Relative Humidity in evening is less than equal to 45.305% then having high pests incursion of aphids (322113).



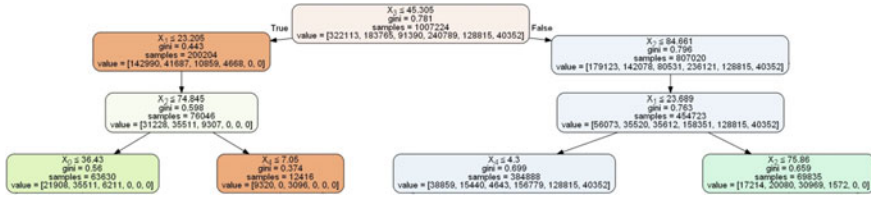


Fig. 1 Decision tree based on the relative humidity

- Rule 2 If the Rule 1 is true and the temperature low value is less than equal to 23.205° then high pests’ intervention of aphids (142990).
- Rule 3 If the relative humidity in morning is less than equal to 74.845% then high pests occurrence of jassids (35511).
- Rule 4 If the maximum temperature is less than equal to 36.43° then there is high pests’ intrusion of jassids (35511).
- Rule 5 If the rainfall is less than equal to 7.05 mm then having high pests invasion of Aphids (9320).
- Rule 6 If the relative humidity in morning is less than equal to 63.435% then having high pests raid of aphids (4672).
- Rule 7 If the minimum temperature is less than equal to 27.2° then having high pests incidence of thrips (3096).
- Rule 8 If the maximum temperature is less than equal to 34.85° then having high pests incidence of jassids (33968).
- Rule 9 If the relative humidity in evening is less than equal to 44.215° then having high pests appearance of aphids (111762).
- Rule 10 If the maximum temperature is less than equal to 29.55° then having high pests occurrence of aphids (6216).
- Rule 11 If the relative humidity in evening is less than equal to 44.495% then having high pests occurrence of aphids (6208).
- Rule 12 If the rainfall is less than equal to 14.5 mm then having high pests attacks of Aphids (105546).
- Rule 13 If the rainfall is less than equal to 37.8 mm then having high pests attacks of aphids (7772).
- Rule 14 If the minimum temperature is less than equal to 15.25° then having high pests attacks of aphids (97774).
- Rule 15 If the Rule 1 is false and the relative humidity in morning is less than equal to 84.661% then high pest attacks of aphids (179123).
- Rule 16 If the maximum temperature is less than equal to 34.935° then having high pests attacks of jassids (81854).
- Rule 17 If the maximum temperature is less than equal to 35.595° then having high pests occurrence of thrips (9288).
- Rule 18 If the maximum temperature is less than equal to 30.415° then having high pests occurrence of jassids (80310).

- Rule 19 If the minimum temperature is less than equal to  $21.62^{\circ}$  then having high pests occurrence of aphids (123050).
- Rule 20 If the minimum temperature is less than equal to  $30.055^{\circ}$  then having high pests occurrence of aphids (79234).
- Rule 21 If the minimum temperature in morning is  $15.62^{\circ}$  then having high pests occurrence of whitefly (27938).
- Rule 22 If the relative humidity in morning is less than equal to 91.41% then having high pests occurrence of aphids (65238).
- Rule 23 If the minimum temperature is less than equal to  $23.689^{\circ}$  then having high pests occurrence of whitefly (158351).
- Rule 24 If the rainfall is less than equal to 4.3 mm then having high pests attacks of whitefly (156779).
- Rule 25 If the relative humidity in evening is less than equal to 75.86% then having high pests attacks of thrips (30969).
- Rule 26 If the maximum temperature is less than equal to  $36.785^{\circ}$  then having high pests attacks of jassids (10808).
- Rule 27 If relative humidity in morning is less than equal to 79.205% then having high pests attacks of thrips (24773).
- Rule 28 If the relative humidity in morning is less than equal to 74.993% then having high pests attacks of whitefly (96227).
- Rule 29 If the relative humidity in evening is less than equal to 66.643% then having high pests attacks of whitefly (60552).
- Rule 30 If the minimum temperature is less than equal to  $23.6295^{\circ}$  then the pest average is less than equal to 0.04% and relative humidity in morning is less than equal to 84.6614%.
- Rule 31 If the maximum temperature is less than equal to 34.385 then the pest average is less than equal to 0.14%.

In all the rules generated, it has been observed that attack of aphids, jassids, thrips is more in almost all weather conditions, out of which aphids is considered as the highest attacking pest.

## 6 Conclusion

In this paper, cotton pests' classification is analyzed using machine learning algorithm. To analyze this big data set and do the classification of pests based on weather, machine learning is helpful for agriculture sector. Maximum temperature, minimum temperature, relative humidity in morning and relative humidity in evening, rainfall are the circumstantial features which influence the pests. Analysis of the present dataset is done by applying three algorithms names K-NN, Naive Bayes and Decision tree. The final result proved that application of decision tree algorithm of machine learning has given the best result in analyzing the classification of pests of cotton crop in various climatic conditions.

## References

1. N. El-Wakeil, A. Abdallah, Cotton pests and the actual strategies for their management control, in *Cotton: Cultivation, Varieties and Uses* (Nova Science Publishers, 2012). ISBN: 978-1-61942-746-4
2. M. Kumar, M. Nagar, *Big Data Analytics in Agriculture and Distribution Channel* (IEEE, 2017). ISBN: 978-1-5090-4890-8
3. V.P. Gandhi, D. Jain, *Cotton Cultivation in Andhra Pradesh, Introduction to Biotechnology in India's Agriculture* (2016), pp. 73–84
4. P. Revathi, M. Hemalatha, Classification of cotton leaf spot diseases using image processing edge detection techniques, in *2012 International Conference on Emerging Trends in Science, Engineering and Technology* (2012). <https://doi.org/10.1109/incoset.2012.6513900>
5. J. Patil, V.D. Mytri, A prediction model for population dynamics of cotton pest using multilayer—perceptron neural network. *Int. J. Comput. Appl.* **67**(4) (2013)
6. D. Sagar, R.A. Balikai, Insecticide resistance in cotton leafhopper, AMRASCA Bigguttula, Bigguttula (ISHIDA)—a review. *Biochem. Cell. Arch.* **14**(2) (2014). ISSN 0972-5075
7. A. Dey, D. Bhoomik, K.N. Dey, Automatic detection of whitefly pest using statistical feature extraction and image classification methods. *Int. Res. J. Eng. Technol.* (2016). e-ISSN: 2395-0056
8. P. Rajan, B. Radhakrishnan, A survey on different image processing techniques for pest identification and plant disease detection. *Int. J. Comput. Sci. Netw.* **5**(1) (2016)
9. A. Das, A.K. Dey, Leaf disease detection, quantification and classification using digital image processing. *Int. J. Innov. Res. Electr. Electron. Instrum. Control Eng.* **5**(11) (2017)
10. M.H. Javed, M. Humair Noor, B.Y. Khan, N. Noor, T. Arshad, K-means based automatic pests detection and classification for pesticides spraying. *Int. J. Adv. Comput. Sci. Appl.* **8**(11) (2017)
11. F.Y. Osisanwo, J.E.T. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, J. Akinjobi, Supervised machine learning algorithms: classification and comparison. *Int. J. Comput. Trends Technol.* **48** (2017)
12. A. Badage, Crop disease detection using machine learning: Indian agriculture. *Int. Res. J. Eng. Technol.* **05**(09) (2018)
13. M. Akila, P. Deepan, Detection and classification of plant diseases by using deep learning algorithm. *Int. J. Eng. Res. Technol.* (2018). ISSN 2278-0181 ICONNECT
14. K. Kitikidou, N. Arambatzis, Big data analysis in agriculture and forestry: a bibliography review. *Res. J. For.* **9**(1), 15 (2015)
15. J. Vidushi, A. Upadhyay, Mongo DB and NoSQL databases. *Int. J. Comput. Appl.* (0975–8887) **167**(10) (2017)
16. B. Gupta, A. Rawat, A. Jain, A. Arora, N. Dhama, Analysis of various decision tree algorithms for classification in data mining. *Int. J. Comput. Appl.* **163**(8) (2017)

# Effect of Formulation Variables on Optimization of Gastroretentive In Situ Rafts of Bosentan Monohydrate HCl by 3<sup>2</sup> Factorial Design



B. Sarada, G. Srividya, R. V. Suresh Kumar, M. Keerthana, and M. Vidyavathi

**Abstract** *Objectives* The study was focused to prepare gastro retentive in situ rafts a controlled release dosage form of Bosentan monohydrate hydrochloride (BMH) through which the administration by oral route becomes easier and retain the drug more time in the stomach, at its absorption window. *Materials and Methods* Nine (F1–F9) gastro retentive in situ rafts of BMH were formulated using 3<sup>2</sup> factorial design by taking ratio of polymers, sodium alginate and pectin (X<sub>1</sub>) and quantity of effervescent (X<sub>2</sub>) as two variables at 3 levels and characterized to find out the influence of two variables on Gel strength (Y<sub>1</sub>), Floating time (Y<sub>2</sub>), Floating lag time (Y<sub>3</sub>), in vitro drug release (Y<sub>4</sub>), Viscosity, Gelling time and Drug content. Then, the results of all responses (Y<sub>1</sub>–Y<sub>4</sub>) were fit into mathematical model as per the design. *Results* Release kinetic studies revealed that, F6 formulation shown first order release as the best fit model. Polymers sodium alginate and pectin were found to possess good compatibility with BMH without any mutual interaction as per DSC and FTIR spectra. The best selected formulation F6 has shown in vitro sustained drug release up to 12 h which was also confirmed by in vivo X-ray image studies in rabbits. *Conclusion* The best selected formulation F6 was found to process in vitro extended release up to 12 h which was also confirmed by in vivo X-ray image studies in rabbits. The stability studies, revealed that, there was no noticeable variation in drug content, floating ability during stability studies.

**Keywords** Gastroretentive in situ rafts · 3<sup>2</sup> factorial design · Bosentan monohydrate HCl (BMH) · Geriatric patients

---

B. Sarada · G. Srividya · M. Keerthana · M. Vidyavathi (✉)  
Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [vidyasur@rediffmail.com](mailto:vidyasur@rediffmail.com)

R. V. Suresh Kumar  
Department of Surgery and Radiology, S. V. Veterinary University, Tirupati, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_5](https://doi.org/10.1007/978-3-030-46939-9_5)

## 1 Introduction

Oral route of administration is the most acceptable and frequently used method for drug delivery. The conventional oral dosage forms of drugs with narrow absorption window in the gastrointestinal tract (GIT) are known to have poor bioavailability due to fast emptying and incomplete drug release at the site of absorption [1]. Gastro retentive drug delivery systems (GRDDS) is the best alternative for controlled delivery of such drugs by slow release of drug for a extended period of time continuously [2, 3]. GRDDSs are appropriate for drugs which have narrow absorption window, drugs which are targeted to gastrointestinal tract, drugs which are sensitive to intestinal fluids [4, 5].

In situ raft systems are in the liquid form at room temperature but, when they come in contact with body fluids or if any change in pH, undergo gelation [6, 7]. These are easier to swallow in liquid form than in strong gel form. The in situ raft system forms a gel which is lighter than fluids in gastric system hence floats on fluid of the stomach and produces gastric retention of dosage form results in prolonged drug delivery in the gastrointestinal tract.

BMH is a dual endothelin receptor antagonist used in the treatment of Pulmonary Arterial Hypertension (PAH), which is a progressive disease indicated by inflated pulmonary arterial pressure (PAP) and pulmonary vascular resistance (PVR), proceeds to right ventricular failure and death [8]. BMH has poor bio-availability, low solubility and poor absorption at intestinal pH. Thus gastro retentive in situ raft of Bosentan monohydrate was selected for self administration of controlled release dosage form easily by oral route to geriatric patients and to maintain the drug for a long time in blood by retaining dosage form the stomach for more time [9]. Half-life of BMH is 5.4 h [10] so, to decrease the periodicity of dosing and to enhance the amenability of patients, the present study was aimed at development of BMH gastro retentive in situ rafts to maintain the plasma drug concentration for long time by ease of swallowing.

## 2 Experimental

### 2.1 Materials

Bosentan monohydrate HCl was a gift sample from Dr. Reddys laboratories, Hyderabad. Calcium carbonate, Pectin, Sodium alginate (SA), Tri sodium citrate, Calcium chloride and Methyl paraben were purchased from SD fine chem. India.

**Table 1** Layout of  $3^2$  factorial design

Variables	Actual variables at 3 levels		
	-1	0	+1
Ratio of pectin: Sodium alginate ( $X_1$ )	1:0.5	1:1	1:1.5
Quantity of effervescent (mg) ( $X_2$ )	500	750	1000

## 2.2 Optimization of Variables Using $3^2$ Factorial Design

A  $3^2$  factorial design was used in the present study by utilizing the design expert<sup>®</sup> software version 10. In this design, two independent variables, different ratios of polymers (pectin: sodium alginate) ( $X_1$ ) and different concentrations of effervescent ingredient, calcium carbonate ( $X_2$ ) were chosen each at three levels (Table 1) and experimental trials were performed for all nine possible combinations as shown in the Table 2. Gel strength ( $Y_1$ ), Floating time ( $Y_2$ ), Floating lag time ( $Y_3$ ), in vitro drug release ( $Y_4$ ) were determined for all formulations [11, 12].

## 2.3 Method of Preparation

Pectin and sodium alginate polymer solutions were prepared separately according to composition as shown in Table 2 over a mechanical stirrer in distilled water containing tri sodium citrate and mixed. Then methyl paraben was added to avoid microbial contamination and degradation of solutions. Then, known quantity of drug was dispersed in distilled water and it was added to the polymer solutions and stirred thoroughly. Then, the specified quantity of calcium carbonate dispersion was added and stirred. Finally, calcium chloride was added to the above solution and mixed for 10–15 min. The formed solution was sonicated for 10 min and it was stored in amber coloured containers for further evaluation [13].

## 2.4 Methods of Evaluation

The prepared  $F_1$ – $F_9$  formulations were evaluated by following methods.

### Gel Strength ( $Y_1$ )

It was determined using a gel strength apparatus (a modified physical balance). It was measured in triplicate on a modified physical balance. It consisted of 2 pans. Lower surface of one of the pans was adhered to raft (1 g) taken in a petri plate. The weights were added into the other pan [14]. The weight at which the lower surface of pan was detached from raft of petri plate was recorded and the gel strength was calculated by using a formula.



$$\text{Gel strength} = M.g/a$$

M: Weight at which the two surfaces detached g: Gravitational force a: Area of surfaces

### **Floating Ability ( $Y_2$ and $Y_3$ )**

The floating ability of formulations was determined in 0.1 N HCl (pH 1.2) taken in a dissolution vessel. Which was placed at  $37 \pm 0.5$  °C and 10 mL of formulation was added into the vessel of dissolution. The time, the formulation took to emerge on the surface of dissolution medium after formation of gel (floating lag time)  $Y_3$  and the duration of time, the formulation continuously floated on to the top of the dissolution medium (floating time)  $Y_2$  were noted [15].

### **In-Vitro Drug Release ( $Y_4$ )**

The in vitro release of BMH from all formulations was conducted with 0.1 N HCl as dissolution medium using USP dissolution test apparatus type II (paddle method) (Lab India D5 8000) at a rotating speed of 50 rpm for 12 h. Samples were collected at different time points and were analyzed using single beam UV spectrophotometer (Shimadzu UV, 1801) at 272 nm in triplicate. The % drug dissolved was calculated at each time point with the help of standard curve [16].

The results for various responses ( $Y_1$ ,  $Y_2$ ,  $Y_3$ ,  $Y_4$ ) were fit into mathematical model as per the design. The perturbation plots and response surface plots were obtained along with polynomial equations for each dependent variable separately by the *Design Expert* software (version 10).

### **Viscosity**

Viscosity of all the formulations was determined in triplicate with the use of Brookfield digital viscometer (Model: LV DV-E). The fixed volume (20 mL) of formulation was poured into a beaker kept at room temperature and spindle LV-64 was used for determination of viscosity [17].

### **Gelling Time**

The ability of prepared raft formulations to form gel in in vitro was determined by taking 10 mL of the formulation in 100 mL of 0.1 N HCl, when the medium comes in contact with the formulation, it was immediately changed into stiff gel like structure. Then the time taken to form gel was noted as gelling time [18].

## **2.5 Drug Content**

About 10 mL of in situ raft (containing 62.5 mg of BMH) was taken and added into 100 mL of volumetric flask containing 50–70 mL of 0.1 N HCl and shaken on magnetic stirrer for 30 min, then sonicated for 15 min until uniform dispersion of all ingredients and filtered. Then, 0.1 N HCl was added upto 100 mL. From this solution,



10 mL of sample was taken and made to 100 mL with 0.1 N HCl, [19] and the % BMH was calculated using double beam UV-Visible spectrophotometer (UV-1800, Shimadzu) at 272 nm with the help of standard curve.

## 2.6 Overall Desirability (OD) Factor

The OD was used for selection of the best formulation, by combining all the responses in order to get desired characteristics. Optimized gastro retentive in situ raft should have less gelling time, high drug content and high floating time for fast action and prolonged release. The individual desirability of each formulation was calculated using the following method [20–22]. The following formula was used for calculation of desirability for gelling time which was minimized.

$$ID_1 = Y_{\max} - Y_i / Y_{\max} - Y_{\text{target}} \quad (1)$$

$$ID_1 = 1 \text{ for } Y_i < Y_{\text{target}}$$

where  $ID_1$  is the individual desirability of gelling time.

The floating time and drug content values were maximized and the following formula was used for their calculation.

$$ID_2 \ \& \ ID_3 = Y_i - Y_{\min} / Y_{\text{target}} - Y_{\min} \quad (2)$$

$$ID_2 \ \& \ ID_3 = 1 \text{ for } Y_i > Y_{\text{target}}$$

where  $ID_2$  and  $ID_3$  were the individual desirability of floating time and drug content respectively.

$$OD = (ID_1 ID_2 ID_3 \dots ID_n)^{1/n} \quad (3)$$

where  $n$  = number of desirable responses of the experiment.

## 2.7 Drug-Excipient Compatibility Studies

DSC studies were performed with pure drug and the best formulation to understand the behavior of polymers with drug on application of thermal energy. DSC (Mettler star<sup>e</sup> SW 8.10) was performed at a heating rate of 10 °C/min in the temperature range of 0–250 °C.

FT-IR studies were also performed by placing the gel in a sample tube and scanned from 4000 to 400  $\text{cm}^{-1}$  using FT-IR spectrophotometer (Bruker). The possible

interaction of BMH with pectin and sodium alginate was accessed by comparing FTIR spectra of pure drug (BMH), polymers (pectin and sodium alginate) and in situ gel formulation [23].

## 2.8 *In-Vivo Studies*

The in vivo efficacy of the best formulation in biological system was evaluated in rabbits after taking IAEC approval (CPCSEA/1677/PO/Re/S/2012/IAEC/12) using radiographic study. X-ray radio graphs were taken at different time intervals to find the in vivo GI retention behavior of the selected best raft prepared using 20% barium sulphate.

The X-ray photographs of GIT of animals (3 young and healthy male New Zealand white strain rabbits) were taken. The animals were selected to monitor the in vivo transit behavior after ensuring that the animals had no signs or previous history of GI illness. In order to stabilise the conditions of GI motility, the animals were kept in fasting condition for 12 h before starting of the experiment. Initially, the X-ray photographic image of rabbit was recorded to confirm the absence of any of the radio-opaque substance in gastro intestine tract of rabbit. The first radiographic image of the animal was taken to ensure absence of radio-opaque material in the GIT. A specified quantity (10 mL) of the formulation containing barium sulfate was administered through a gastric tube into rabbit under anaesthesia in lateral recumbancy. The presence of the formulation was monitored by X-ray radiographs at different time intervals.

## 2.9 *Short Term Stability Studies*

The stability of in situ raft was estimated after filling and sealing in light protective amber colored bottles with rubber caps and aluminum covering. These were stored at three different temperatures and relative humidity (i.e.  $25 \pm 2$  °C,  $60\% \pm 5$ ;  $30 \pm 2$  °C,  $65\% \pm 5$ ; and  $40 \pm 2$  °C,  $65\% \pm 5$ ) as per ICH guidelines [24] and were inspected visually and evaluated for every 15 days for their in vitro gelling time, floating ability and drug content. Then the results were compared with the results of before storage and stability was assessed.

## 3 Results and Discussion

In the present work, nine floating gastro retentive in situ rafts of BMH (F<sub>1</sub>–F<sub>9</sub>) were prepared as per 3<sup>2</sup> factorial design (Tables 1 and 2) and characterized to find out the effect of two variables i.e., ratio of polymers (X<sub>1</sub>) and quantity of effervescent (X<sub>2</sub>)

in formulation on Gel strength ( $Y_1$ ), Floating time ( $Y_2$ ), Floating lag time ( $Y_3$ ), In vitro drug release ( $Y_4$ ), Viscosity, Gelling time and Drug content. The results of the responses ( $Y_1$ – $Y_4$ ) of all the prepared formulations (F1–F9) are shown in Table 3, the respective perturbation plots (Fig. 1), response surface and contour plots (Fig. 2) and quadratic Eqs. 1–4 were generated.

$$Y_1 = 468.18 + 47.91 X_1 - 14.53 X_2 \quad (1)$$

$$Y_2 = 6.51556 + 2.276667 X_1 - 0.066667 X_2 \quad (2)$$

$$Y_3 = 13.73333 + 1.32000 X_1 - 12.66333 X_2 \\ + 2.56000 X_1 X_2 - 2.760000 X_1^2 + 2.78000 X_2^2. \quad (3)$$

$$Y_4 = 74.99278 - 49.95167 X_1 - 30.29000 X_2 \\ + 0.050000 X_1 X_2 + 27.11333 X_1^2 - 9.14667 X_2^2. \quad (4)$$

### 3.1 Effect of Polymer Ratio ( $X_1$ ) on $Y_1$ – $Y_4$

The coefficient of  $X_2$  was positive indicated positive effect of  $X_1$  on  $Y_1$ ,  $Y_2$ , and  $Y_3$  (Eqs. 1–3) but, it was negative in Eq. 4 indicated negative effect of  $X_1$  on  $Y_4$ . It indicated that as polymer ratio  $X_1$  was increased, gel strength, floating time, floating lag time ( $Y_1$ – $Y_3$ ) were also increased where as ( $Y_4$ ) drug release was extended with improving polymer ratio (Fig. 3) might be due to high retarding effect and increased viscosity at high quantity of polymer. The same results were also observed in its respective perturbation plot (Fig. 1), which indicated that gel strength and floating time were also increased with the increase in ratio of polymers ( $X_1$ ) due to increased quantity of polymer and increased hydrophilic role of sodium alginate in gelation of raft.

### 3.2 Effect of Quantity of Effervescent ( $X_2$ ) on $Y_1$ – $Y_4$

The coefficient of  $X_2$  was negative in all Eqs. 1–4, indicated negative effect of  $X_2$  on  $Y_1$ – $Y_4$ . The gel strength, floating lag time and time of floating were decreased with increase in  $X_2$  i.e. the quantity of effervescent,  $\text{Ca}_2\text{CO}_3$ . The decreased gel strength might be due to increased gas formation and entrapment which led to reduced contacts between polymer molecules and decreased floating lag time ( $Y_2$ ) was observed due to formation of effervescence in gel with decreased density led to fast floating. The negative effect of  $X_2$  on  $Y_3$  might be due to decreased gel strength.

**Table 3** Results of different parameters of prepared nine formulations

Formulation code	X <sub>1</sub>	X <sub>2</sub>	Y <sub>1</sub> (gel strength) N/m <sup>2</sup> Mean ± SD	Y <sub>2</sub> (floating time) h Mean ± SD	Y <sub>3</sub> (floating lag time) min Mean ± SD	Y <sub>4</sub> (drug release) % Mean ± SD	Viscosity (Cps) Mean ± SD	Drug content (%) Mean ± SD
F <sub>1</sub>	-1	-1	463.66 ± 3.05	8.3 ± 0.22	5.0 ± 0.1	73.50 ± 1.05	6260 ± 0	90.1 ± 1.0
F <sub>2</sub>	-1	0	476.66 ± 2.30	7.8 ± 0.08	4.7 ± 0.40	80.42 ± 1.55	7443 ± 5.13	94.8 ± 1.6
F <sub>3</sub>	-1	+1	509.1 ± 2.00	9.7 ± 0.069	2.83 ± 0.15	79.68 ± 1.12	6233 ± 11.54	91 ± 2.4
F <sub>4</sub>	0	-1	473.88 ± 3.60	7.5 ± 0.26	2.9 ± 0.1	77.88 ± 1.24	1465 ± 5.13	96.6 ± 1.4
F <sub>5</sub>	0	0	384.81 ± 4.35	8.1 ± 0.55	2.9 ± 0.26	76.22 ± 2.55	8693 ± 60.27	93.4 ± 1.6
F <sub>6</sub>	0	+1	543.39 ± 1.52	11.2 ± 0.6	3.0 ± 0.11	90.79 ± 1.79	9573.4 ± 1.0	97.72 ± 2.1
F <sub>7</sub>	+1	-1	347.32 ± 3.60	7.4 ± 0.53	2.67 ± 0.05	75.2 ± 0.52	12,566 ± 5.0	90.9 ± 4.1
F <sub>8</sub>	+1	0	459.22 ± 0.77	8.2 ± 0.88	2.44 ± 0.35	83.8 ± 1.01	6954 ± 34	94.54 ± 2.8
F <sub>9</sub>	+1	+1	535.65 ± 2.64	10 ± 0.56	3.13 ± 0.04	75.2 ± 0.98	7663 ± 11.54	92 ± 2.5

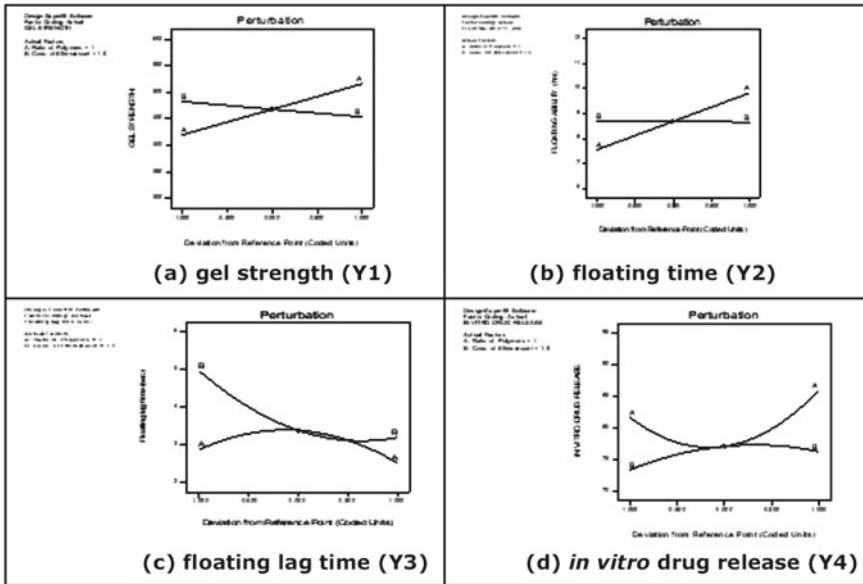


Fig. 1 Perturbation plots of different parameters (Y<sub>1</sub>–Y<sub>4</sub>)

### 3.3 Comparison of Effect of X<sub>1</sub> and X<sub>2</sub>

The value of coefficient of X<sub>1</sub> was higher (47.91 for gel strength, 2.27 for floating time) than the value of coefficient of X<sub>2</sub> (14.53 for gel strength, 0.066 for floating time) indicated that the ratio of polymers (X<sub>1</sub>) has more influence on gel strength and floating time than the quantity of effervescent ingredient (X<sub>2</sub>). X<sub>1</sub> has shown more effect than X<sub>2</sub> on Y<sub>3</sub>, indicated the floating time was majorly influenced by polymers and its quantities but not by the quantity of effervescent. Both variables X<sub>1</sub> and X<sub>2</sub> have shown negative influence on drug release (Y<sub>4</sub>) indicated that increase in quantity of polymer and effervescent, the decreased release suggested the prolonged release due to increased retarding effect by increased thickness or path length of polymer at higher proportions of polymers through which drug diffuses. The influence of polymer ratio (X<sub>1</sub>) was more on drug release than quantity of effervescent (X<sub>2</sub>) as the coefficient of X<sub>1</sub> was greater than the coefficient of X<sub>2</sub>. All the above observations were also found in their respective perturbation (Fig. 1), response surface and counter plots (Fig. 2).

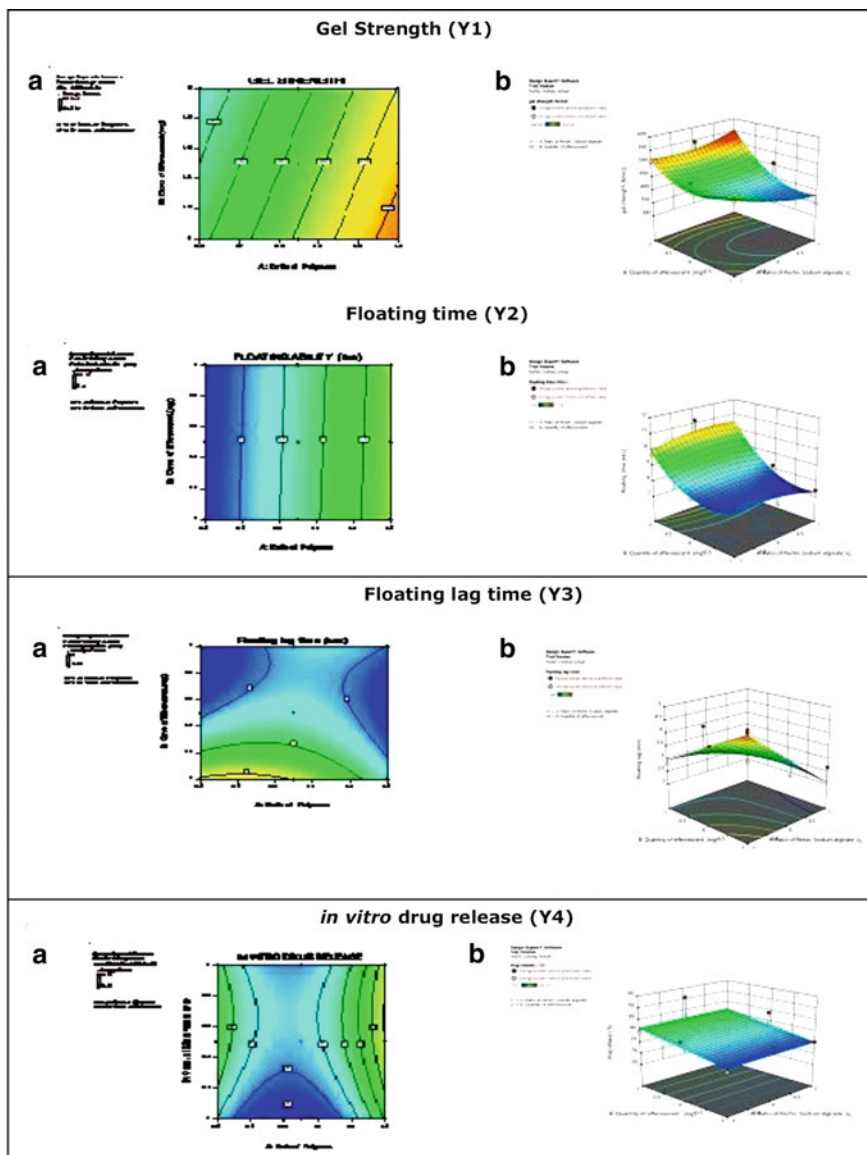


Fig. 2 a Two dimensional contour plots, b three dimensional response surface plots of different parameters (Y<sub>1</sub>–Y<sub>4</sub>)

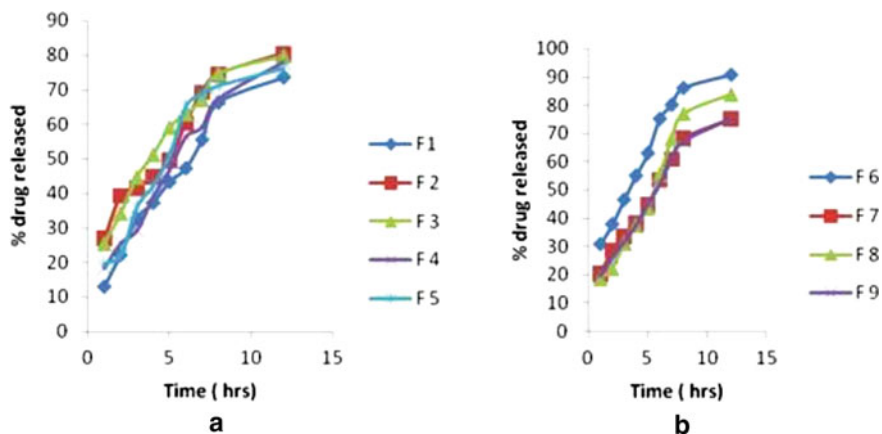


Fig. 3 Dissolution profile of prepared gastro retentive in situ rafts **a** F<sub>1</sub>–F<sub>5</sub>, **b** F<sub>6</sub>–F<sub>9</sub>

### 3.4 Other Parameters

Then the viscosity, gelling time and drug content were determined for all the formulations. As per Table 2, the range of gelling time was 1–3 s for all formulations and there was no significant difference in % drug content of all the formulations. The viscosity was increased with increase in ratio of polymers ( $X_1$ ) due to formation of poly electric complex between pectin and sodium alginate. Using the above results, OD was calculated for all formulations. The OD value of F6 was 0.8752 which was nearer to one among all the nine formulations. Hence, F6 was considered as the best formulation.

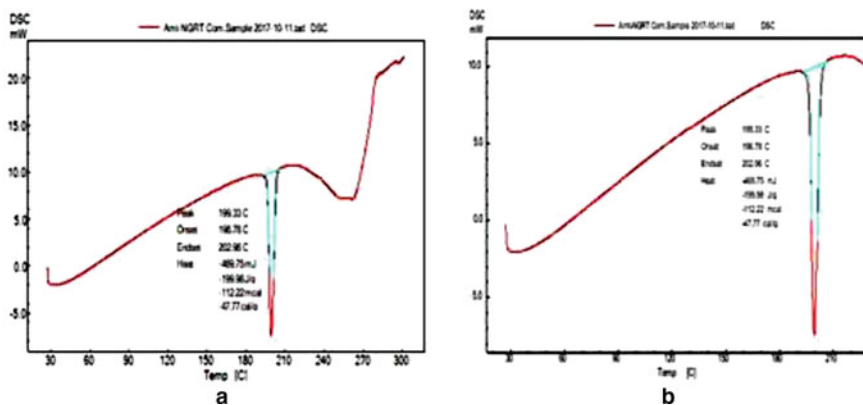
### 3.5 Characterization of the Best Selected Formulation

#### In Vitro Release Kinetics

The mechanism of drug release from F6 was found by subjecting the in vitro release data into release kinetic studies. The regression value was closer to unity in the case of first order plot ( $R^2 = 0.987$ ), so the release mechanism was found to be in first order (Table 4). The data indicated poor linearity and was less than the value of the first order plot, when it was plotted according to the zero order equation. The regression coefficient obtained for Higuchi model was found to be superior in comparison with Korsmeyer peppas plot, which indicated that the drug release was first order controlled by Fickian diffusion.

**Table 4** Kinetics of In vitro release from best formulation

Model	R <sup>2</sup>	N
Zero order	0.9337	—
First order	0.987	—
Higuchi	0.891	—
Korsmeyer Peppas	0.829	1.02

**Fig. 4** DSC thermograms of **a** BMH, **b** best in situ raft

### 3.6 Compatibility Studies

The compatibility of drug with polymers was determined through DSC and FTIR analysis. DSC thermo grams of pure drug and formulation F6 have shown same endothermic peak at 199.3 °C (Fig. 4), at the melting point of drug indicated the drug was compatible with other excipients present in the formulation.

FTIR spectroscopy analysis of pectin, sodium alginate, pure Bosentan monohydrate HCl, and F6 was conducted. The spectra of pure drug and in situ raft (F6) demonstrated the characteristic absorption peaks at 2749.52  $\text{cm}^{-1}$  for O–H stretching, 2964.59  $\text{cm}^{-1}$  for C=O, 3341.86  $\text{cm}^{-1}$  for N–H, 1354.28  $\text{cm}^{-1}$  for C–C, 1575.56  $\text{cm}^{-1}$  for C–H, 1205.38  $\text{cm}^{-1}$  for C–N indicated that the drug characteristics were not changed after developing a formulation.

### 3.7 In Vivo Studies

The radio photo graphs of GIT of rabbits at different time periods after oral administration of F6 in Fig. 5 clearly indicated that the raft was retained successfully in the stomach up to 12 h. The increased gastric retention time of the formulation was



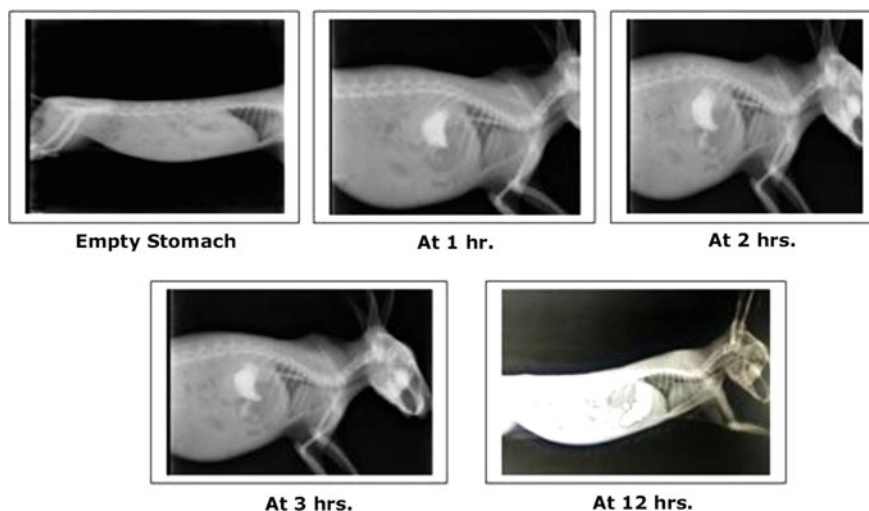


Fig. 5 In vivo X-ray images showing gastric retention

due to the floating ability and gel strength of combined polymers. It confirmed the residence of prepared raft in in vivo for 12 h.

### 3.8 Stability Studies

Short time stability studies were carried out for the best selected formulation (F6) for 3 months. The samples were characterized periodically for every 15 days, and found that there were no changes in gelling time, drug content and floating ability upon storage (Table 5) confirmed the stability of prepared formulation.

## 4 Conclusion

The present work confirmed the successful, stable formulation of bosentan monohydrate hydrochloride gastro retentive in situ raft with sodium alginate and pectin for prolonged release upto 12 h to treat pulmonary arterial hypertension in particular with ease of swallowing by geriatric patients.

**Table 5** Results of different parameters of optimized gastro retentive in situ raft under stability studies

Days	% drug content				Gelling time (min)			
	25 °C ± 2 °C 60% ± 5RHMean ± S.D.	30 °C ± 2 °C 60% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.	25 °C ± 2 °C 60% ± 5RHMean ± S.D.	30 °C ± 2 °C 60% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.
15	96.62 ± 0.01	97.18 ± 0.27	95.15 ± 0.03		2.0 ± 0.01	2.09 ± 0.003	2.14 ± 0.004	
30	97.15 ± 0.02	98.0 ± 0.3	97.68 ± 0.04		1.9 ± 0.008	2.00 ± 0.008	2.17 ± 0.05	
60	96.92 ± 0.7	96.9 ± 0.004	96.18 ± 0.12		1.6 ± 0.005	1.5 ± 0.003	1.7 ± 0.003	
90	95.27 ± 0.12	95.17 ± 0.10	97.01 ± 0.015		1.8 ± 0.005	1.6 ± 0.004	1.7 ± 0.001	
Days	Floating lag time (min)				Floating time (h)			
	25 °C ± 2 °C 60% ± 5RHMean ± S.D.	30 °C ± 2 °C 60% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.	25 °C ± 2 °C 60% ± 5RHMean ± S.D.	30 °C ± 2 °C 60% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.	40 °C ± 2 °C 65% ± 5RHMean ± S.D.
15	8.14 ± 0.12	8.74 ± 0.06	9.13 ± 0.004		12.01 ± 0.14	12.30 ± 0.04	11.50 ± 0.7	
30	9.0 ± 0.09	8.12 ± 0.005	10.16 ± 0.003		11.3 ± 0.05	12.40 ± 0.1	12.00 ± 0.7	
60	10 ± 0.011	9.9 ± 0.017	11.0 ± 0.12		11.14 ± 0.7	11.01 ± 0.7	11.58 ± 0.5	
90	9.17 ± 0.004	9.5 ± 0.007	10.17 ± 0.18		0.012 ± 0.7	12.05 ± 0.1	120 ± 0.5	

## References

1. A.A. Despande, C.T. Rhodes, N.H. Shah, Controlled release drug delivery system for prolonged gastric residence: an overview. *Drug Dev. Ind. Pharm.* **22**(6), 531–539 (2008)
2. U. Kumar, B. Mandal, F.G.S. Chatterjee, Gastro-retentive drug delivery systems and their in-vivo success: a recent update. *Asian J. Pharm.* **11**(5), 575–584 (2016)
3. N. Brahma, K. Singh, H.K. Won, Floating drug delivery systems: an approach to oral controlled drug delivery via gastric retention. *J. Control Release.* **63**(3), 235–259 (2000)
4. S. Vaibhav, M. Rakesh, N. Tanaji, Development and optimization of a floating multiparticulate drug delivery system for norfloxacin. *Turk. J. Pharm. Sci.* **16**(3), 326–334 (2019)
5. U.K. Mandal, B. Chatterjee, F.G. Senjoti, Gastro-retentive drug delivery systems and their in-vivo success: a recent update. *Asian J. Pharma. Sci.* **11**, 575–584 (2016)
6. S.P. Vyas, R.K. Khar: Gastro retentive systems, in *Controlled Drug Delivery*, 6th edn, (Vallabh Prakashan, 2006), pp 197–217
7. R. Hetangi, V. Patel, M. Moin, In situ gel as a novel approach of gastro retentive drug delivery. *Int. J. Pharm. Life Sci.* **1**(8), 440–447 (2010)
8. S.A. Said, M.E.M. Abdalla, Role of endogenous endothelin-1 in stress-induced gastric mucosal damage and acid secretion in rats. *Regul. Pept.* **73**(1), 43–50 (1998)
9. Ch.S. Vijayvani, M. Vidyavathi, Preparation and in vitro characterization of Bosentan mono hydrate mucoadhesive microsphere. *Eur. J. Pharm. Med Res.* **3**(5), 340–350 (2016)
10. R. Harish, T.E.G.K. Murthy, K.B. Chandrasekhar, Formulation and evaluation of bosentan solid dispersion. *Asian J. Pharm.* **11**(1), 75–81 (2017)
11. J. Lou, W. Hu, R. Tian, H. Zhang, Y. Jia, J. Zhang, L. Zhang, Optimization and evaluation of a thermoresponsive ophthalmic insitu gel containing curcumin-loaded albumin nanoparticles. *Int. J. Nano med.* **9**, 2517–2525 (2014)
12. A. Sujata, P. Sradhanjali, R.P. Nihar, Optimization of HPMC and carbopol concentrations in non-effervescent floating tablet through factorial design. *Carbohydr. Polym.* **102**, 360–368 (2014)
13. R.A. Shah, M.R. Mehta, D.M. Patel, C.N. Patel, Design and optimization of mucoadhesive nasal in situ gel containing sodium cromoglycate using factorial design. *Asian J. Pharm.* **5**(2), 65–74 (2011)
14. A. Gupta, S. Garg, R.K. Khar, Measurement of bioadhesion strength of mucoadhesive buccal tablet: design of an in vitro assembly. *Indian Drugs* **30**, 152–155 (1992)
15. R.J. Rishad, C.N. Patel, M.P. Dashrath, N.P. Jivani, Development of a novel floating in-situ gelling system for stomach specific drug delivery of the narrow absorption window drug baclofen. *Iran. J. Pharm. Res.* **9**(4), 359–368 (2010)
16. H. Gupta, S. Jain, R. Mathur, A.K. Mishra, T. Vepandian, Sustained ocular drug delivery from a temperature and pH triggered novel in situ gel system. *Drug Deliv.* **14**(8), 507–515 (2007)
17. W. Kubo, S. Miyazaki, D. Attwood, Oral sustained delivery of paracetamol from in situ gelling gellan and sodium alginate formulation. *Int. J. Pharm.* **258**(1–2), 55–64 (2003)
18. R.P. Patel, B. Dadhani, R. Ladani, A.H. Baria, J. Patel, Formulation, evaluation and optimization of stomach specific in situ gel of Clarithromycin and Metronidazole benzoate. *Int. J. Drug Deliv.* **2**, 141–153 (2010)
19. W. Vinay, M. Mohan Varma, S. Manjunath, Formulation and evaluation of stomach specific in-situ gel of metoclopramide using natural, bio-degradable polymers. *Int. J. Res. Pharma. Biomed. Sci.* **2**(1), 193–201 (2011)
20. K.P. Bhavin, H.P. Rajesh, S.A. Pooja: Development of Oral sustained release rifampicin loaded chitosan nanoparticles by design of experiment. *J. Drug Deliv.* 1–10 (2013)
21. R.C. Mashru, V.B. Sutariya, M.G. Sankalia, J.M. Sankalia, Effect of pH on in vitro permeation of ondansetron hydrochloride across porcine buccal mucosa. *Pharm. Dev. Technol.* **10**(2), 241–247 (2005)
22. P.G. Paterakis, E.S. Korakianiti, P.P. Dallas, D.M. Rekkas, Evaluation and simultaneous optimization of some pellets characteristics using a 3<sup>3</sup> factorial design and the desirability function. *Int. J. Pharm.* **248**(1–2), 51–60 (2002)

23. C. Renu, B. Swati, Drug-Excipient compatibility screening-role of thermoanalytical and spectroscopic techniques. *J. Pharm. Biomed. Anal.* **87**, 82–97 (2014)
24. B. Sanjay, S. Dinesh, S. Neha, Stability testing of pharmaceutical products. *J. Appl. Pharma. Sci.* **02**(03), 129–138 (2012)

# Performance Analysis of Apache Spark MLlib Clustering on Batch Data Stored in Cassandra



K. Anusha and K. UshaRani

**Abstract** With the tremendous increase in the amount of data being generated from variety of sources there is a need of efficient data storage and processing techniques. Some of the sources generating this large amount of data are Weather Sensors, Scientific experiments, etc. This huge voluminous data is termed as BigData. Due to ever-increasing amount of data there is a demand for faster data ingestion and processing. Apache Spark, a dominant processing tool is a publicly available platform for processing outsized data and is mostly intended for iterative machine learning jobs. In this study, an integrated approach i.e., Spark MLlib Clustering on batch weather data stored in Cassandra database is proposed. This helps to analyze our data into number of Clusters which is required and useful for further examination of data. The main idea of this study is to evaluate Batch Processing performance of an integrated approach with two popular clustering algorithms.

**Keywords** Big data · Apache Spark MLlib · Clustering · Apache Cassandra

## 1 Introduction

Now-a-days distinct types of data are growing speedily in extent, complication and some data processing techniques need to be used for further analysis. Apache Spark is a public domain processing tool of great use which is fault-resilient and multipurpose cluster engine.

Spark's Machine Learning Library (Spark MLlib), one of the advanced and high level libraries of Spark-Core is a publicly known machine learning library which store and operate on data with benefits of data-parallelism [1]. Spark MLlib offers a variety of machine-learning algorithms like classification, regression, clustering and

---

K. Anusha (✉) · K. UshaRani  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [siri.bachina@gmail.com](mailto:siri.bachina@gmail.com)

K. UshaRani  
e-mail: [usharanikuruba@yahoo.co.in](mailto:usharanikuruba@yahoo.co.in)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_6](https://doi.org/10.1007/978-3-030-46939-9_6)

so on. To quickly run machine learning algorithms in an iterative method without compromising its performance it utilizes the Apache Spark's distributed computing architecture.

Clustering is nothing more than assembling certain objects in such a manner where items in a similar set termed cluster is further comparable with items in remaining clusters. This can be the core activity of exploratory data analysis which is general method for statistical analysis of data utilized in several areas together with machine learning [2]. Clustering is an all-purpose activity which requires to be solved. Distinct clustering algorithms are available which differ significantly in establishing clusters and effectiveness of finding them.

Clustering is a most powerful method which is principally utilized in numerous forecasting jobs like weather forecasting, storm detection, and so on. Clustering procedure is probably helpful to evaluate data effectively and produce crucial information. Originating even and constant clusters via clustering methods enhances data accurateness and efficiency.

In this study, weather data is considered for experimentation. The weather data is divided into distinct clusters based on temperature and precipitation on the basis of minimizing the within sum of squared errors (WSSE) between objects and centroids which is a useful metric to choose best number of clusters. This procedure continues repeatedly until it no longer creates an alteration. The most popular and well known Clustering algorithms K-Means and Bisecting K-Means are considered.

K-means, most popular clustering method is well acknowledged for its simplicity with less time complexity. K-Means segregates the data pixels into predetermined number of groups [3].

Another popular Clustering Algorithm is Bisecting K-Means algorithm which is a modification over basic K-Means algorithm. It mainly based on K-Means algorithm and keeps the merits of K-Means along with having some advantages over K-Means.

Hence, to get the benefits of popular clustering techniques the proposed integrated approach i.e., Spark MLlib and Cassandra with Clustering is proposed. The proposed method is experimented by conducting two experiments with K-Means as well as with Bisecting K-Means to divide the experimented batch weather data into clusters which is helpful for future prediction and also to evaluate the efficiency of batch processing based on time and WSSE of data clusters. The results are compared and analyzed to suggest the best integrated method for Batch Processing of weather data with optimum clusters i.e., clusters with minimum WSSE.

The rest of the paper is organized as follows: Sect. 2 describes Literature Review; Sect. 3 represents the Methodology; Sect. 4 represents the Proposed System; Sect. 5 represents the experimental analysis; Sect. 6 represents the conclusion.

## 2 Literature Review

Meng et al. [1] proposed that Spark was essentially computationally efficient and therefore suitable for large-scale machine learning applications. They introduce

Apache Spark's popular library for machine learning, MLlib. The library benefits from parallelism of data to accumulate and function on data.

Ghosh et al. [4] introduced a new SMS spam identification method that mainly based on Apache Spark MLlib as platform. They explained that there are many research papers available using Weka for SMS spam classification but Apache Spark MLlib is completely new in this race. Various machine learning algorithms such as logistic regression with L-BFGS, NB, DT and gradient boosted trees will be used for comparison.

Gnanaraj et al. [5] proposed a new K-Means algorithm for retrieving hidden knowledge by forming a cluster of unified structures and datasets.

Harifi et al. [6] presented a summary of Spark MLlib algorithms. The clustering methods like Power-Iteration, Gaussian Mixture Model and so on are fully described.

Abirami et al. [7] shows the comparison of different clustering algorithms of segmentation model and determines which algorithm is best for the user. They used a number of key stages such as preprocessing, pattern discovery and pattern analysis, similarity measurement and clustering technique to improve efficiency.

Assefi et al. [8] explained that they have implemented different experiments on machine learning to investigate quality and quantity of system. They also highlight present directions in research of big data and offer ideas of upcoming effort.

Chaudhari et al. [9] presented a scalable and new method named "Smart Cassandra and Spark Integration" to address challenges in integration of Cassandra and Spark for system management. For evaluating the performance, SCSI Streaming framework is compared.

Jayanthi et al. [10] proposed implementation of Spark for analyzing weather data consider various weather stations and make view of different parameters like mean precipitation, hot or cold, etc.

### 3 Methodology

Brief description of Spark MLlib, Cassandra and Clustering models are presented in this section.

#### 3.1 *Apache Spark Machine Learning Library*

Apache Spark MLlib, an adaptable library of machine learning from Apache foundation. It is an open source framework which helps to communicate among several ML algorithms. This library supports various languages like Scala, Python, R and Java. The primary API of machine learning for Apache Spark is Data Frame based API. The Data Frame API for Spark MLlib offers an identical API over Machine Learning methods and over several dialects. It offers many regular learning methods for instance Classification, Regression, Clustering, etc.

### 3.2 Clustering Models

Clustering algorithms are mainly used for extracting knowledge. Clustering is mainly used for grouping of similar data to form clusters [5]. K-Means technique is a majorly utilized method for cluster examination. Bisecting K-means is variation of K-means method and is a divisive hierarchical clustering algorithm. These two Clustering models are experimented on weather data for performance evaluation of an integrated approach. The results are compared and analyzed to suggest the best integrated framework for Batch Processing.

#### K-Means Clustering

It is a popular technique for partitioning a dataset into 'k' groups automatically [SEM17]. In this k is supposed to be unaltered or permanent. Allow 'k' prototypes ( $w_1, w_2 \dots w_k$ ) be initialized to any of 'n' input samples ( $i_1, i_2 \dots i_n$ ). As a result,  $w_j = i_l$  where  $j \in \{1, 2 \dots k\}$ ,  $l \in \{1, 2 \dots n\}$ .  $C_j$  is jth group and its value is disjoint split of input samples [11].

Cluster centers and data points in every one cluster are adjusted using an iterative algorithm to finally classify k clusters [11].

#### Steps to identify k Clusters

1. Arrange 'K' pixels in the gap portrayed by items that are being clustered. These points represent initial group centroid.
2. Allot every item to cluster that have nearest centroid.
3. At the point after every object is allotted, location of k centroids is recomputed.
4. Rehash 2 and 3 steps till centroids never again progress. This creates a segregation of items to clusters and standard to be diminished is determined.

#### Bisecting K-Means Clustering

This is a conjunction of two clustering methods K-Means and hierarchical. As an alternative of splitting chosen data to k number of groups in all iterations, this method partitions single group into two sub clusters in every bisecting phase until required k groups is acquired [7].

#### Steps to identify k clusters

1. Choose one group to partition.
2. Discover two splits utilizing essential K-Means method.
3. Reiterate 2nd step i.e., bisecting step for ITER times and obtain partition which creates groups with complete resemblance.
4. Rehash all steps till ideal amount of groups are gained.

Since it depends on K-Means, along with benefits of K-Means it has a minimum of interest above K-Means. It is increasingly productive with enormous size of k [7].



### 3.3 *Apache Cassandra*

NoSQL tools provide various procedures incorporated in storing and retrieving of data other than tables operated in conventional tools. They support multiple copies of data to avoid data loss, comprise of simpler API, ultimately reliable, handles voluminous data. Apache Cassandra, a NoSQL Database, is highly scalable distributed tool designed toward handle huge data volumes on distinct storage machines that provide large accessibility without fail [12].

### 3.4 *Spark Cassandra Integration*

The traditional databases are not able to handle large datasets because of their limited capacity. NoSQL databases are one of the solutions to store Big Data, which overcome the problem of traditional relational databases. Apache Cassandra, one of the NoSQL databases is a publicly available distributed database management system. Cassandra is an extendable non relational database that provides high accessibility, performance, etc. Cassandra's design has the capability to scale, perform and present uninterrupted uptime. Cassandra has various key features and benefits and is a vital database for current online use cases [13].

In our previous work we proposed Spark-Cassandra integration and Spark SQL-Cassandra integration for Batch Processing to evaluate the performance of both frameworks on batch data [13, 14]. The results of two integrated methods are compared and hence observed that Spark SQL Cassandra Integration has better Processing Performance on batch data.

To further analyze the performance of Big Data Frameworks over Batch Processing, in this study a new integrated method i.e., Spark MLlib and Cassandra with Clustering is proposed.

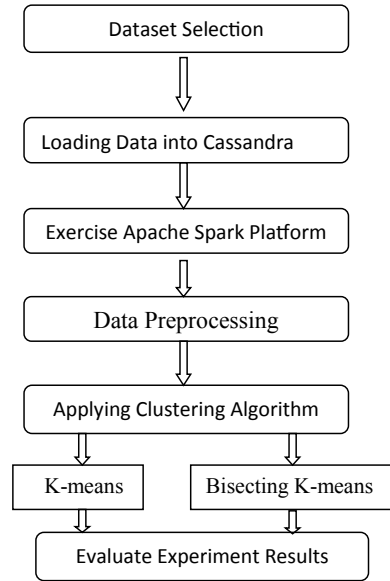
## 4 Proposed System

An integrated method, Spark MLlib and Cassandra is proposed. Two Spark MLlib Clustering algorithms are experimented on weather data stored in Cassandra. The sequence of steps in this method is.

### 4.1 *Proposed Algorithm*

1. Load Weather data CSV file into Cassandra database.
2. Start Spark-Cassandra Shell.

**Fig. 1** Sequence of steps for proposed method



3. Connect to Cassandra CQL shell.
4. Read data from Cassandra using Spark SQL.
5. Apply Clustering on Cassandra data using Spark MLlib.
6. Measure the required parameters.

The complete process for Batch Processing the data using the proposed integrated method is shown in the following Fig. 1.

## 4.2 Dataset Selection

In this study, we considered weather data for experimentation. Dataset is gathered from National Climatic Data Center (NCDC). The weather data consists of attributes such as date, location id and observations for each weather parameters like temperature, moisture, pressure and so on.

## 5 Experimental Results

Proposed integrated method is tested on K-Means and Bisecting K-Means methods on Cassandra data to measure the time taken to split data into multiple clusters and also to calculate the WSSE which indicates the goodness of a cluster. A high WSSE

suggests that the items in a cluster have differences and may not be useful. These parameters helps to identify the best proved clustering algorithm on proposed method for analysis of data.

**Experiment 1: Spark MLlib—Cassandra Integrated Method with K-Means**

In first experiment, the proposed approach with K-Means is experimented to retrieve the batch data stored in Cassandra by dividing data into different clusters.

Distinct cluster sizes are considered. From clusters 1 to 5 time decreases but error rate is varying. And from clusters 6–9 time remains constant and error rate also goes down. Again from cluster 10, time remains constant but error rate is increasing drastically. Hence, cluster 9 is considered as best.

The performance of proposed integrated method with K-Means is measured by using time efficiency and WSSE. The results are presented in the following Table 1 and the best cluster is highlighted.

From the above table it can be observed that time and WSSE are minimum for cluster 9. So, time and WSSE values of proposed approach with K-Means with 9 clusters give better performance with minimum WSSE and time.

The results are illustrated graphically in the following Fig. 2.

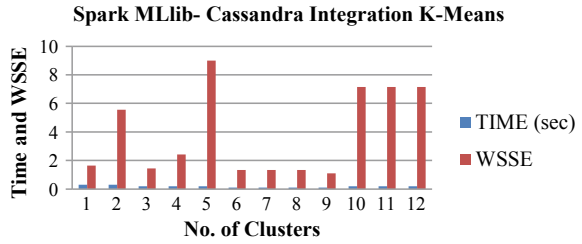
**Experiment 2: Spark MLlib—Cassandra Integrated Method with Bisecting K-Means**

In second experiment, proposed approach with Bisecting K-Means is experimented to retrieve the batch data stored in Cassandra by dividing data into different clusters. As in the first experiment, distinct clusters are considered and repeatedly perform the execution to find the optimum cluster with minimum WSSE and time. In this experiment, from clusters 1 to 4 time decreases and error rate varies slightly but at cluster 5 there is an extreme increase in error rate. However, again from cluster 5 the

**Table 1** Performance of integrated approach with K-Means

No. of clusters	Spark MLlib—Cassandra integration with K-Means	
	Time (s)	WSSE
1	0.3	1.64315
2	0.3	5.55240
3	0.2	1.43875
4	0.2	2.42818
5	0.2	8.99238
6	0.1	1.34495
7	0.1	1.33755
8	0.1	1.32952
<b>9</b>	<b>0.1</b>	<b>1.09486</b>
10	0.2	7.1539
11	0.2	7.15500
12	0.2	7.15500

**Fig. 2** Performance of integrated approach with K-Means



time remains almost constant as well as error rate also goes down and it is observed that the error rate at cluster 10 decreases to 1.57621 with processing time of 0.4 s. So, we considered experimenting further clusters to check whether there is a further decrease in WSSE value. But the error rate increases drastically for cluster sizes 11 and 12. Hence, cluster 12 is considered as the stopping criteria.

The performance of proposed integrated method with Bisecting K-Means is measured by using time efficiency and error rate. The results are presented in the following Table 2 and the best cluster is highlighted.

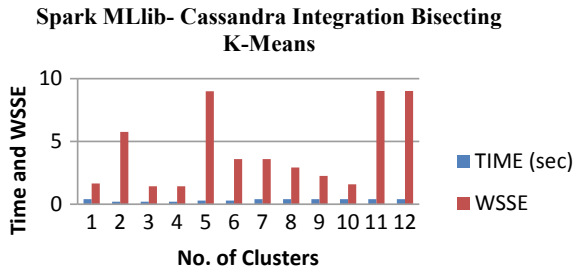
From the above table it can be observed that time and error rate is minimum for cluster 4. Even though the error rate at cluster 10 decreases to 1.57621 with processing time of 0.4 s, cluster 4 is considered as optimum cluster as it has minimum WSSE value of 1.43848 with 0.2 s of processing time which is better compared to cluster 10. So, time and WSSE values of proposed approach with Bisecting K-Means with 4 clusters of batch weather data gives better performance with minimum WSSE and time.

The results are illustrated graphically in following Fig. 3.

**Table 2** Performance of integrated approach with bisecting K-Means

No. of clusters	Spark MLlib—Cassandra integration bisecting K-Means	
	Time (s)	WSSE
1	0.4	1.64315
2	0.2	5.75240
3	0.2	1.43875
<b>4</b>	<b>0.2</b>	<b>1.43848</b>
5	0.3	8.99238
6	0.3	3.59877
7	0.4	3.59877
8	0.4	2.92432
9	0.4	2.25026
10	0.4	1.57621
11	0.4	9.01244
12	0.4	9.01244

**Fig. 3** Performance of integrated approach with bisecting K-Means



### 5.1 Comparison of Proposed Integrated Approach with Two Clustering Techniques

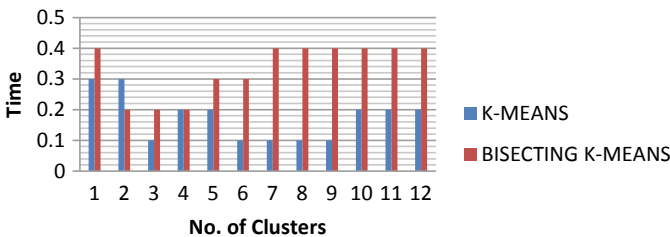
In this study the benchmarking was done on the comparison of two experiments on proposed integrated approach i.e., Spark MLlib and Cassandra with K-Means and Bisecting K-Means on weather data to find out optimum cluster size with minimum WSSE and time. The experimental results are shown in Tables 1 and 2 related to all the clusters. In both experiments the best cluster with minimum SSE and time is considered for comparison and is illustrated in the following Table 3.

From the above table it is very clear that the proposed approach with K-Means with 9 clusters is having optimum time and WSSE compared to proposed method with Bisecting K-Means. Hence, the proposed approach with K-Means with 9 clusters is recommended for future prediction of weather data.

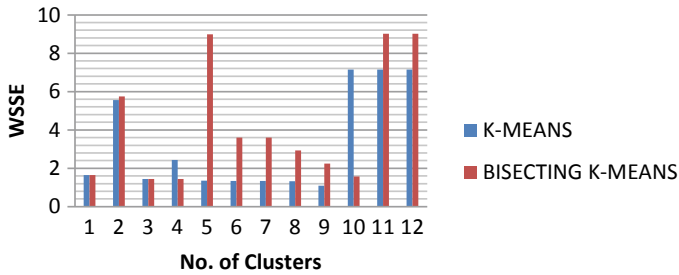
The comparison of time efficiency of integrated approach with K-Means and Bisecting K-Means experiments are illustrated in graphical representation in following Fig. 4.

**Table 3** Comparison of time efficiency and WSSE of the integrated approach

Integrated approach with clustering techniques	No. of clusters	Time (s)	WSSE
Spark MLlib-Cassandra integration with K-Means	9	0.1	1.09486
Spark MLlib-Cassandra integration with bisecting K-Means	4	0.2	1.43848



**Fig. 4** Comparison of time efficiency of integrated approach



**Fig. 5** Comparison of WSSE of integrated approach

The comparison of error rate of integrated approach with K-Means and Bisecting K-Means experiments are illustrated in graphical representation in following Fig. 5.

## 6 Conclusion

To get the benefits of popular clustering techniques to partition the weather data which is required for further analysis an integration of Spark Machine Learning Library (Spark MLlib) and Cassandra with two popular Clustering algorithms like K-Means and Bisecting K-Means is proposed on weather data and its performance is evaluated based on the time and WSSE.. The results proved that proposed integrated approach with K-Means is effective in terms of time and minimum Sum of Squared Errors (SSE). Hence, best integrated method i.e., Spark MLlib and Cassandra with K-Means is recommended for future prediction of weather data.

## References

1. X. Meng, MLlib: machine learning in Apache Spark. *J. Mach. Learn. Res.* **17** (2016)
2. [https://en.wikipedia.org/wiki/Cluster\\_analysis](https://en.wikipedia.org/wiki/Cluster_analysis)
3. <https://spark.apache.org/docs/latest/mllib-clustering.html>
4. A. Ghosh, A. Kumar Pasayat, Identifying spam SMS using Apache Spark Mllib. *J. Emerg. Technol. Innov. Res.* **5**(5) (2018). ISSN: 2349-5162
5. T. Nelson Gnanaraj, K. Ramesh Kumar, N. Monica, Survey on mining clusters using new k-mean algorithm from structured and unstructured data. *Int. J. Adv. Comput. Sci. Technol.* **3**(2) (2014). ISSN: 2320-2602
6. S. Harifi, E. Byagowi, M. Khalilian, *Comparative Study of Apache Spark MLlib Clustering Algorithms Conference Paper* (2017)
7. K. Abirami, P. Mayilvahanan, Performance analysis of K-means and bisecting K-means algorithms in Weblog data. *Int. J. Emerg. Technol. Eng. Res. (IJETER)* **4**(8) (2016)
8. M. Assefi, E. Behraves, G. Liu, A.P. Tafti, Big data machine learning using Apache Spark MLlib, in *Conference IEEE Big Data 2017*, Boston, USA (2017)
9. A. Chaudhari, P. Mulay, *SCSI: Real-Time Data Analysis with Cassandra and Spark Research Gate* (2019)

10. D. Jayanthi, G. Sumathi, Weather data analysis using spark—an in-memory computing framework, in *International Conference on Innovations in Power and Advanced Computing Technologies* (2017)
11. <https://dzone.com/articles/cluster-analysis-using-apache-spark-exploring-colo>
12. [https://www.tutorialspoint.com/cassandra/cassandra\\_introduction.htm](https://www.tutorialspoint.com/cassandra/cassandra_introduction.htm)
13. K. Anusha, K. UshaRani, Big data techniques for efficient storage and processing of weather data. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* 5(VII) (2017). ISSN: 2321-9653
14. K. Anusha, K. Usha, Rani performance evaluation of Spark SQL for batch processing, in *Advances in Intelligent Systems and Computing*. Accepted for publication in Springer series

# A Study on Opinion of B.Sc. Nursing Students on Health Informatics and EMR as Part of Nursing Education



B. GangaBhavani

**Abstract** In India nursing educational institutions run under the umbrella of Indian Nursing council, which prescribes the curriculum and other norms for nursing education and nursing care services in health care sector. Indian nursing council has prescribed Introduction of computers in first year of B.Sc. nursing course and in second year, B.Sc. nursing course Communication and Education Technology. **Objectives:** The present study aims at knowing the nursing graduate student's knowledge about health informatics and EMR and their opinion about these courses to be included in nursing curriculum. **Methodology:** An incidental study is conducted on final year B.Sc. nursing students. **Analysis and Results:** Eighty percent of the male students are in favor of EMR system and seventy percent of the female students are in favor of it. Seventy eight percent of the both male and female students want to upgrade the nursing curriculum by including EMR and Health Informatics in the syllabus. Overall seventy four percent of the students are of the opinion that nursing curriculum need to be updated giving scope for technology based elective subjects like Electronic Medical Records and Health Informatics.

**Keywords** Knowledge · Opinion · Nursing students · Health informatics · Electronic medical records · Nursing curriculum · Electives subjects

## 1 Introduction

Nursing is one of the key services of health care system. The genesis of nursing is as old as motherhood, but the systematic study of nursing was started by Florence Nightingale. In India nursing educational institutions run under the umbrella of Indian Nursing council, which prescribes the curriculum and other norms for nursing education and nursing care services in health care sector. Computers have entered in every educational system including health care institutions. Hence Indian nursing council has included a paper on Introduction of computer in first year of

---

B. GangaBhavani (✉)  
College of Nursing, SPMVV, Tirupati, India  
e-mail: [gangabhavani259@gmail.com](mailto:gangabhavani259@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_7](https://doi.org/10.1007/978-3-030-46939-9_7)



B.Sc. nursing course with 15 h of theory and 30 h of practical instruction which is of a very basic, in nature. In 2nd year B.Sc. nursing course Communication and Education Technology with 60 h of theory and 30 h of practical work is prescribed for study. Both the subjects are related to the use of technology for health care communication. Global technical advancements resulted in the emerging new courses as Health Informatics and Electronic Medical Records/Electronic Health Records.

According to Wikipedia “Health Informatics (also called health care informatics, medical informatics, nursing informatics, clinical informatics or biomedical informatics) is information Engineering applied to the field of health care, essentially the management and use of patient health care information” [1] ([https://en.wikipedia.org/wiki/Health\\_informatics](https://en.wikipedia.org/wiki/Health_informatics)).

Electronic medical record is “An electric (digital) collection of medical information about a person that is stored on a computer. An electronic medical record includes about a patient’s health history such as diagnosis, medicine, tests, allergies, immunizations and treatment plans. Electronic medical records can be seen by all health care providers who are taking care of the patient” [2] (Fig. 1).

EMR chart

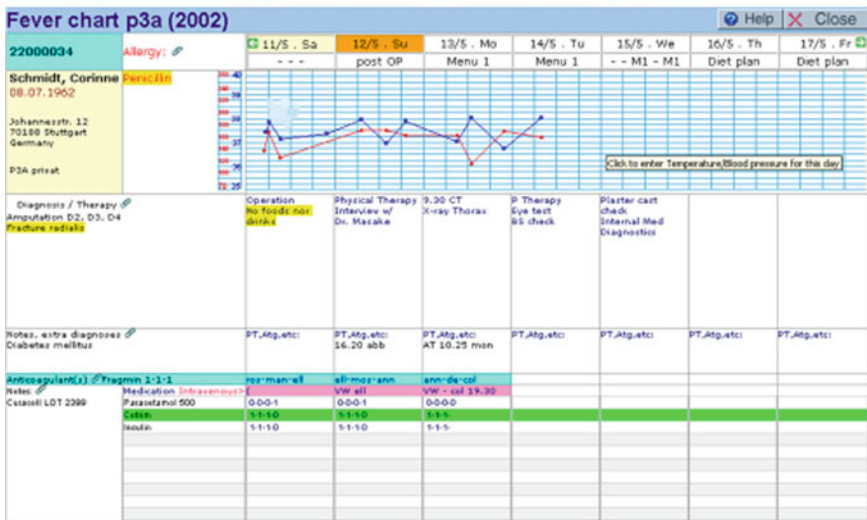


Fig. 1 Courtesy [to.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-medical-record](https://www.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-medical-record)

## 2 Need for the Study

Present era is for the students who are technology savvy, nursing students are no exception to it. Computers and internet had reached even the remote rural areas of India.

The present study aims at knowing the nursing graduate students' opinion about the use of technology in patient care and communication. Their opinion about health informatics and EMR system to be included in nursing curriculum.

## 3 Objectives of the Study

- To find out the knowledge of B.Sc. nursing students regarding Health Informatics.
- To know the knowledge of B.Sc. nursing students regarding EMR system.
- To find out the opinion of B.Sc. nursing students regarding Health Informatics and EMR system to be included in nursing curriculum as elective subjects.

## 4 Review of Literature

Trained Nurses' Association and Student Nurses Association of India has conducted National level conference with the theme "Empowering Nurses Through Advanced Technology" on 17th and 18th August 2017 in Dr. D. Krishna Murthy Kalakshetram, Tirupati, Andhra Pradesh.

The review of the literature is taken from the Souvenir published after the conference.

A survey was conducted on, "Attitude of Nurses Regarding use of Computers And Electronic Patient Record" by D. Beula, K. Manjula, M. S. Anusha. According to the study nurses working in public sector hospitals of Andhra Pradesh are positive towards Electronic Medical Records system [3]. Same findings are reported in a study conducted by the investigator on the public sector nurses of Mumbai.

"A study to Assess the knowledge, perception and attitude on EMR among nurses in selected hospitals at Tirupati" by G. S. Dilli Rani revealed that there is a significant positive relation between staff nurses' level of education and knowledge scores on EMR system [4].

S. Spandanahals conducted "A study to Assess Nursing Faculty's Knowledge on Importance of Nursing Informatics". The study revealed a significant association between the nursing informatics' knowledge and computer literacy of the nursing faculty [5].

"A study on Effectiveness of Structured Teaching program on Knowledge and Utilization of Computers Among Nursing Staff in Gandhi Hospital, Secunderabad,"

is done by J. C. KavithaLatha. The study results revealed that there is a significant improvement in use of computers in patient care documentation among nurses after the training program.

## 5 Methodology

An incidental study is conducted on final year B.Sc. nursing students of different nursing colleges at Tirupati when they are attending a cultural fest arranged by a leading nursing college of Tirupati on 28th September 2019. Total number of samples is 100.

Inclusion criteria is all those students who are studying 4th year B.Sc. nursing course and about to be graduated within 3 months. Exclusion criteria is those students who are unwilling to participate in the study.

Informed consent is taken from the students.

## 6 Analysis of the Data

**Total n = 100** Male students are 20 and female students are 80.

Results of the Data analysis (Table 1).

### 6.1 Knowledge Scores on EMR and Health Informatics

- Seventy percent of the students felt that using computers to document patient care improves their self-image.
- Ninety six percent of the students feel satisfied if they document patient care accurately in computers.
- Sixty percent of the students have the knowledge that Health Informatics system helps in Patient billing and insurance claims.

**Table 1** Knowledge and opinion scores of the B.Sc. nursing students on health informatics and EMR

Categories	Female (%)	Male (%)	Grand total (%)
Knowledge scores on H.I.	67	74	71
Knowledge scores on EMR	77	79	78
Opinion scores on H.I. and EMR	76	80	78

- Seventy one percent of the students are aware about the basic qualifications to study Health Informatics courses.
- Sixty seven percent of the students have the knowledge regarding the principles of Health informatics system.
- Seventy two percent of the students know the scope of EMR.
- Sixty nine percent of the students are aware of the benefits and uses of EMR system.
- Seventy percentage of the students know the limitations and demerits of EMR.
- Sixty-two percentage of the students know how to recognize the technical error.
- Seventy-five percentage of the students are aware that EMR system is time saving.

## ***6.2 Analysis of the Opinion Scores***

- Sixty-eight percentage of the students are confident of using computers for recording the patient care while thirty-two percentage opted for paper pen documentation.
- Eighty percent of the students are of the opinion that EMR is helpful for patients' referrals.
- Ninety percent of the students are of the opinion that if they make mistakes in EMR Patient will be affected drastically.
- Ninety eight percent of the students are of the opinion that technology is not an enemy to the health care system. They are in favor of using technology for health.
- Ninety six percent of the students are of the opinion that proper use of EMR and Informatics reduces the medico legal delays.
- Sixty-five percentage of students are of the opinion that EMR can be protected by proper passwords and security keys.
- Ninety percent of the students are apprehensive about patient data leakage and cyber threats.
- Eighty-two percentage of the students want to include EMR in Nursing curriculum.
- Seventy-nine percent of the student want Health Informatics in nursing curriculum as elective subject.
- Ninety percent of the students are of the opinion that introduction to computers as the nursing subject in under graduate nursing courses must be updated.

## ***6.3 Knowledge Scores Based on Gender***

- Male students' knowledge score on EMR is slightly higher as seventy six percent against female students' seventy two percent.
- Overall knowledge score of the total students is seventy-four percentage on EMR.

- Male students' Knowledge score on Health Informatics is higher as seventy-four Against seventy percent of female students.
- Overall knowledge score of the total students is seventy-two percentage on Health Informatics.
- The knowledge scores of both EMR and Health Informatics uses in nursing is seventy three percent.

### 6.4 Opinion Scores Analysis

- Eighty percent of the male students are in favor of EMR system as against seventy percent of the female students.
- Seventy percent of the both male and female students want to upgrade the nursing curriculum by including EMR in the syllabus.
- Seventy two percent of the male students want Health informatics as elective subject in nursing curriculum as against seventy percent of female students.
- Overall ninety-six percent of the students are of the opinion that computers are helpful for nursing care.
- Overall seventy percent of the students are of the opinion that nursing curriculum need to be updated giving scope for technology based elective subjects (Fig. 2).

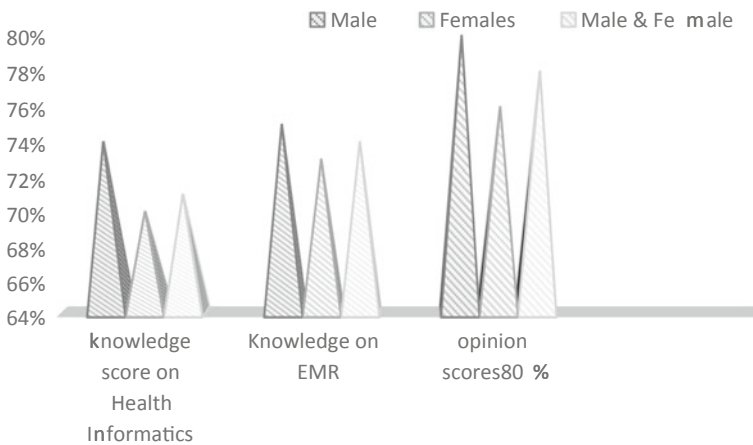


Fig. 2 Knowledge and opinion scores on H.I. and EMR

## 7 Summary and Conclusion

Tirupati as a city of knowledge hub has most of premier educational institutions. Nursing students of the city has the unique opportunity to interact with many students who are technology savvy and adopt the same zeal. This is clearly reflective of the high knowledge scores and well-set opinion about use of technology in nursing care. Indian Nursing Council is going to update the nursing curriculum in 2020. Hopefully technology-based subjects will be included in the curriculum.

## 8 Recommendations

- Same study can be carried out in other cities to know the students' opinion about Electronic Medical Records and Health Informatics.
- Online MOOCS courses in Health Informatics and Electronic Medical Records can be provided on SWAYAM platform for nursing, paramedical and medical students.

## References

1. [https://en.wikipedia.org/wiki/Health\\_informatics.2.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-medical-record](https://en.wikipedia.org/wiki/Health_informatics.2.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-medical-record)
2. D. Beulah, K. Manjula, G. Anusha, Attitude of nurses regarding use of computers and electronic patient record: a survey, Souvenir, in *XXVII TNAI/SNA Biennial State Conference*, Tirupati (2017), pp. 51–52
3. G.S. Dilli Rani, A study to assess the knowledge, perception and attitude on EMR among nurses in selected hospital at Tirupati, Souvenir, in *XXVII TNAI/SNA Biennial State Conference*, Tirupati (2017), pp. 54–55
4. S. Spandana, A study to assess the effectiveness of structured teaching program on level of knowledge regarding the importance of nursing informatics among nursing faculties of selected nursing colleges at Tirupati, Souvenir, in *XXVII TNAI/SNA Biennial State Conference*, Tirupati (2017), pp. 58–59
5. J.C. Kavitha Latha, A study to assess the effectiveness of structured teaching program on knowledge and utilization of computers among nursing staff in Gandhi Hospital, Secunderabad, Souvenir, in *XXVII TNAI/SNA BIENNIAL STATE Conference*, Tirupati (2017), pp. 58–59

# A Comprehensive Hybrid Ensemble Method with Feature Selection Techniques



G. Sujatha and K. Usha Rani

**Abstract** Data Mining is the process of examining huge pre-existing databases in order to produce new information. Decision Tree is a very popular and practical approach in Data Mining. Decision Trees plays a crucial role in medical field. Ensemble methods create multiple models and produce more accurate solutions than a single model. Feature Selection improves the standard of the data by removing irrelevant attributes, due to that accuracy will increase. In this study experiments are conducted on a Hybrid Method i.e., C4.5 Decision Tree with MultiBoostAB Ensemble technique with different Feature Selection Techniques on Tumor datasets for finding out accuracy, execution time to build a decision tree and size of the tree.

**Keywords** Data mining · C4.5 · MultiBoostAB · Feature selection process · Tumor datasets

## 1 Introduction

In today's information age, there is a requirement for a powerful analytical solution for the extraction of the useful information from the massive quantity of data within the databases. Data Mining is the solution and is used for extracting knowledge from a data set and converts it into the human understandable format. There is more than one form of knowledge extraction from the datasets. These forms are used for extract models connecting vital categories or discover out future knowledge trends [1]. There are two forms in Data Mining: Classification and Prediction. By using the Classification set of models are identified and data classes are distinguished [2]. Decision Trees (DT) are widely used in classification [3–6]. Ensemble classifiers will give best result than a single classifier by considering a preference of their estimated values [7].

---

G. Sujatha (✉) · K. Usha Rani  
Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India  
e-mail: [sujatha.g47@gmail.com](mailto:sujatha.g47@gmail.com)

K. Usha Rani  
e-mail: [usharanikuruba@yahoo.co.in](mailto:usharanikuruba@yahoo.co.in)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_8](https://doi.org/10.1007/978-3-030-46939-9_8)

In Data Mining, Data Preprocessing is used to delete the unnecessary data due to that quality of the data is improved. Data Reduction is one step during the Preprocessing. During the data reduction, Feature Selection or Attribute Selection is crucial task. By using Feature Selection a subset is generated from its original features. Over the years, Feature Selection plays active role in research and development [8]. Feature Selection used for find out the necessary attributes for Data Mining and also used for diminishing the number of irrelevant attributes in the Dataset.

From our previous study, it was proved that the Hybrid method i.e., C4.5 Decision Tree classifier with MultiBoostAB ensemble technique is best for Tumor Datasets [9]. Further, it was proved that committee size of 10 for Primary Tumor (PT) and committee size of 20 for Colon Tumor (CT) having higher accuracy by using the Hybrid Method [10]. A majority of studies have concluded that there is no single feature selection is best method for all the datasets [11, 12]. Hence, for verification of the best suitable method(s) for PT and CT Datasets experiments are conducted with the Hybrid Method by considering all feature selection methods for improving accuracy.

The paper is organized into 4 sections. Section 2 deals with literature survey based on Feature Selection Processes, Sect. 3 deals with the description of different Feature Selection Processes, Sect. 4 consists of Experimental Results with explanation and Concluded by Sect. 5.

## 2 Related Work

Chen et al. [13] carried out experiments with Feature Selection techniques on ensemble methods AdaBoost, MultiBoostAB and Bagging with C4.5 algorithm on medical data. The results are evaluated based on accuracy and execution time. BalaKumar et al. [14] done experiments on E-mail spam classification. In this 3 Feature Selection Methods ReliefF, Cfssubseteqval and Chi square are used with different classification algorithms.

Novakovic [15] identified the impact of Feature Selection with MultiBoostAB + Support Vector Machine (SVM) on 5 medical data sets. These results are compared by using different feature selection methods.

Rokach et al. [16] had done experiments on different datasets by using various Feature Selection methods with C4.5 Decision tree classifier.

Souza et al. [17] analyzed a framework with various Feature Selection techniques LVF, Relief, Focus and Relieved algorithms are classify with various classification algorithms like C4.5, Naive Bayes and K-nearest neighborhood on 13 datasets.

Polat et al. [18] developed a novel algorithm with the name Feature Selection-Artificial Immune Recognition System [FSAIRS] on medical data set by using C4.5 decision classifier.

Deisy et al. [19] carried out experiments on medical data with different Feature Selection methods with C4.5 algorithm.



Hall et al. [20] developed experiments on different data sets with one Feature Selection Technique (Correlation Based Feature Selection) by using C4.5 classifier.

Aruna et al. [21] developed a novel algorithm CSSFFS for Feature Selection on the purpose of detecting Breast cancer. This is a novel and genetic algorithm with the merging of filters and wrappers. Here Support Vector Machines (SVM) acts as a filter and SFFS acts a wrapper in this algorithm.

Lavanya et al. [22] experimented with Feature selection techniques by using CART classifier on different Breast Cancer Datasets. From the results it is observed that the best Feature Selection method depends on particular dataset.

The literature survey clearly shows that there is no single Feature Selection is suitable for all the datasets.

### 3 Background

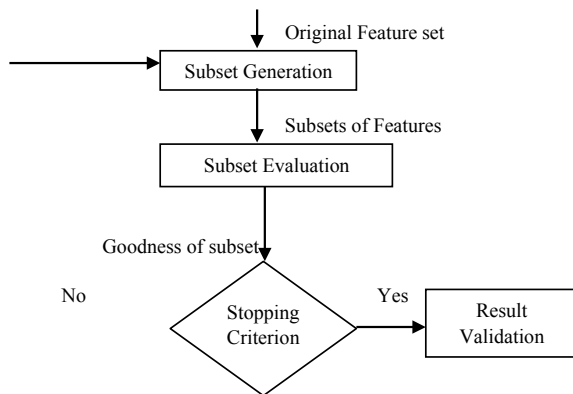
This section describes Feature Selection Process and Feature Selection Techniques which are related to this study.

#### 3.1 Feature Selection Process

The procedure for Feature Selection Process is divided into four steps, which are represented in Fig. 1 [23]. Steps are: Generation of Subset, Subset Evaluation, Stopping Criteria and Result Validation.

- The Subset Generation is used to form the different subsets with various features [23]. This is two types:

Fig. 1 Feature selection process



- Forward Selection: It initializes with empty set and options are added one by one.
- Backward Elimination: It initialized with full set and the options are removed [23, 24].
- Later, new feature set is generated by using the prior best feature set. If the new subset is superior to the prior best feature set, then the existing feature subset is replaced with the new best feature subset.
- This procedure is repeated until it satisfies the Stopping Criterion.
- Get the optimal resultant set based on evaluation criterion.

### 3.2 Feature Selection Methods

Feature Selection Methods are 2 types: Subset Evaluation (SE) and Attribute Evaluation (AE). Further each one is having various methods as represented in the following Fig. 2.

**Subset Evaluation Methods:** A brief description about all the methods under subset Evaluation is presented here:

**Correlation based Feature Selection.** This is feature subset selection algorithm. This works based on correlation among the features. It selects best feature subset by using heuristic evaluation function [24]. It is used for nominal or categorical features.

**Classifier SE.** It evaluates feature subsets with either training data or testing data sets. It estimates the accuracy based on these subsets.

**Filtered SE.** In this learning algorithm is used for evaluation of subsets. For finding out the accuracy cross validation is used in this algorithm [23].

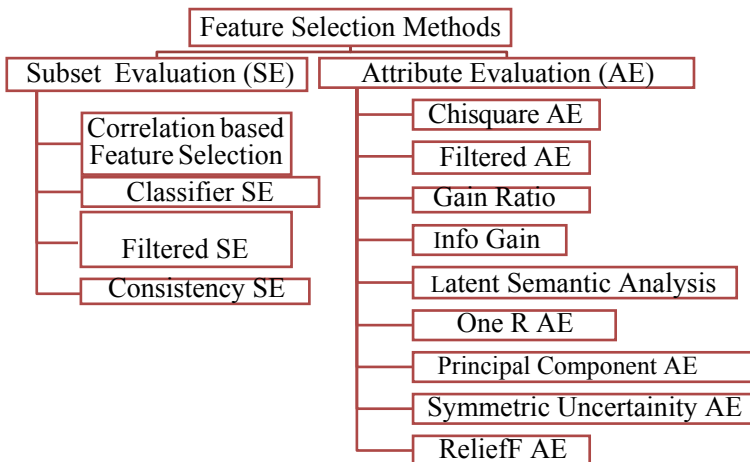


Fig. 2 Methods in feature selection process

**Consistency SE.** In this degree of consistency is used to calculate the feature subsets. Here instead of using full subset it searches for the smallest subset which gives same consistency like the full set.

**Attribute Evaluation Methods:** A brief description about most common methods under Attribute Evaluation is prescribed here:

**Chi-Square AE.** It is used to test whether the class label not depend of a particular feature or not [23]. It is represented as

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

where

*r* Different values for a feature

*c* Number of classes

*O<sub>ij</sub>* Number of instances with *i* value in class *j*

*E<sub>ij</sub>* Number of instances (Expected) with *i* value in class *j*.

**Gain Ratio (GR).** It is used to eliminate the bias in IG. The formula for GR is represented as

$$GR = \frac{Gai(A)}{SplitInfo(A)} \quad (2)$$

Split Info is

$$Split\ Info_A(D) = - \sum_{j=1}^v \frac{(|D_j|)}{(|D|)} \log^2 \frac{(|D_j|)}{(|D|)} \quad (3)$$

where *D* is a set containing of *d* data samples with *n* various classes.

**Information Gain (IG).** It is developed by Hunt in 1996 for determining the feature significance between the attribute and class label based on entropy [24].

$$IG(X, Y) = H(X) - H((X)/(Y)) \quad (4)$$

In this *X*, *Y* are attributes and *H(X)*, *H(Y)* are entropy of *X* and *Y*.

**OneR AE.** These algorithms follow various methods to generate compact, easy to interpret, and exact rules based on a particular class at a time [24]. For the use of decision trees classification rules are used.

$$r = (a, c) \quad (5)$$

where *a* is series of tests that can be evaluated and

*c* is a class that applies to instances covered by ruler.

Rule based algorithm follow basically three steps. They are

1. Generate rule R on datasets
2. Discard the training data which is covered by the rule
3. Do the above 2 steps up to generate one level tree.

**Symmetric Uncertainty (SU).** Correlation based feature selection is the base for Symmetrical Uncertainty (SU) [23].

$$SU = 2.0 \frac{*IG(X,)}{H(X) + H(Y)} \tag{6}$$

SU is in the range of [0, 1] and H(X) and H(Y) are entropies of X and Y.

**ReliefF.** It was proposed by Kira and Rendel [24]. ReliefF is an extension to relief algorithm. It supports dependent features and as well as noisy data. It works by measuring the ability of an attribute in separating similar instances. It follows three basic steps for the ranking of the features. Those are:

1. Calculate the hit and miss values to the nearest rank
2. Calculate feature weight
3. Return top K features based on given threshold.

## 4 Experimental Result

The Experimental Datasets Primary Tumor (PT) is collected from UCI Machine Learning Repository [25] and Colon Tumor (CT) is collected from Bioinformatics Group Seville [26], which are openly accessible. The following Table 1 shows the collected Datasets Information.

If the data contains irrelevant data like missing values or empty cell entries, those cells are preprocessed. In this study, to preprocess the irrelevant data with the corresponding mean of the cell entries is substituted. It was proved that Hybrid Method i.e., c4.5 + MultiBoostAB with the committee sizes of 10 for PT and 20 for CT having high accuracy [9, 10] by using Weka tool on the data using 10-fold cross validation.

To study the performance of the Hybrid Method with various Feature Selection Method(s) various experiments are conducted and the results are shown in the

**Table 1** Datasets information

Data set	Primary tumor (PT)	Colon tumor (CT)
Attributes	18	2001
Classes	2	2
Instances	339	62
Missing values	Yes	No

**Table 2** CfsSubsetEval + MultiBoostAB + C4.5

Search technique	PT				CT			
	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size
Best-first	11	40.41	0.22	66	26	79.03	22.02	7
Exhaustive	11	40.41	1.39	66	–	–	–	–
Genetic	13	40.71	0.27	87	532	75.81	13.52	5
Greedy-step wise	11	40.12	0.06	66	26	75.81	16.5	5
Linear forward selection	11	40.41	0.11	66	<b>21</b>	<b>80.64</b>	<b>0.95</b>	<b>5</b>
Random	<b>11</b>	<b>41.89</b>	<b>0.56</b>	<b>66</b>	–	–	–	–
Rank	<b>11</b>	<b>41.89</b>	<b>0.11</b>	<b>81</b>	<b>47</b>	<b>80.64</b>	<b>43.64</b>	<b>5</b>
Scatter	11	40.12	0.13	66	–	–	–	–
Subset size forward selection	11	40.18	0.36	66	<b>21</b>	<b>80.64</b>	<b>1.2</b>	<b>5</b>

Bold indicates The best feature selection method with corresponding search technique

Tables 2, 3, 4, 5 and 6. The best Feature Selection Method with corresponding Search Technique is highlighted in each Table.

From the Table 2 it is observed that Random and Rank search techniques for PT and Liner Forward Selection, Rank and Subset size Forward Selection search techniques for CT has highest and same accuracy than other search Techniques. For both the Datasets Rank Search Technique is having highest accuracy.

From the Table 3 it is observed that Linear Forward Selection Technique for PT and Random Search Technique for CT have highest accuracy than other techniques.

From the Table 4 it is observed that Scatter Search Technique for PT and Rank Search Technique for CT have highest accuracy than other techniques.

From the Table 5 it is observed that Genetic Search Technique for PT and Subset size Forward Selection Search Technique for CT has highest accuracy than other techniques.

From the Table 6 it is clearly observed that Chi Squared Attribute Eval, Gain Ratio Attribute Eval and Symmetric Uncert Attribute Eval for PT and for CT Gain Ratio Attribute Eval and Symmetric Uncert Attribute Eval are the best and having same accuracy than other Feature Selection Methods.

The best search technique corresponding to a particular feature selection method and from other feature selection methods for the two Tumor Datasets is represented in the Tables 7 and 8.

**Table 3** ClassifierSubsetEval + MultiBoostAB + C4.5

Search technique	PT				CT			
	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size
Best-first	18	24.78	0.19	–	–	64.52	13.89	1
Exhaustive	1	24.78	0.17	1	1	61.29	0.16	3
Genetic	18	24.78	0.02	18	–	64.52	1.42	–
Greedy-step wise	18	24.78	0.09	18	–	64.52	1.5	–
Linear forward selection	<b>8</b>	<b>37.46</b>	<b>31.13</b>	<b>19</b>	–	–	–	–
Random	1	28.02	0.02	3	<b>1</b>	<b>83.87</b>	<b>6.64</b>	–
Rank	–	24.78	0.11	–	–	–	–	–
Scatter	–	24.78	0.09	–	–	64.52	1.52	–
Subset size forward selection	18	24.78	0.19	–	–	64.52	13.89	1

Bold indicates The best feature selection method with corresponding search technique

**Table 4** ConsistencySubsetEval + MultiBoostAB + C4.5

Search technique	PT				CT			
	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size
Best-first	16	43.36	0.36	85	5	77.42	1.84	7
Exhaustive	16	42.48	33.53	85	–	–	–	–
Genetic	16	41.29	0.28	85	205	80.65	0.25	11
Greedy-step wise	16	43.36	0.31	85	5	77.42	0.92	7
Linear forward selection	16	43.36	0.34	85	6	77.42	0.5	7
Random	16	43.36	8.63	85	–	–	–	–
Rank	17	43.07	0.22	85	<b>15</b>	<b>88.71</b>	<b>5.23</b>	<b>7</b>
Scatter	<b>17</b>	<b>43.66</b>	<b>0.67</b>	<b>85</b>	–	–	–	–
Subset size forward selection	16	43.36	0.28	85	3	79.03	0.67	7

Bold indicates The best feature selection method with corresponding search technique

**Table 5** FilteredSubsetEval + MultiBoostAB + C4.5

Search technique	PT				CT			
	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree Size	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size
Best-first	11	39.23	0.03	65	26	80.65	22	7
Exhaustive	11	39.23	0.44	65	–	–	–	–
Genetic	<b>12</b>	<b>41.89</b>	<b>0.09</b>	<b>91</b>	532	75.81	13.54	7
Greedy-step wise	11	39.82	0.03	65	26	77.42	16.28	7
Linear forward selection	11	39.23	0.02	65	19	80.65	0.64	7
Random	11	39.82	0.17	65	–	–	–	–
Rank	12	40.71	0.03	52	–	–	–	–
Scatter	11	39.23	0.03	65	–	–	–	–
Subset size forward selection	11	39.82	0.11	65	<b>19</b>	<b>82.26</b>	<b>1.03</b>	<b>7</b>

Bold indicates The best feature selection method with corresponding search technique

From the Table 7 it is clear that Consistency Subset Eval, Chi Squared Attribute Eval, Gain Ratio Attribute Eval and Symmetric Uncert Attribute Eval feature selection techniques are best for PT dataset because these are having same and highest accuracy.

From the Table 8 it is clear that Consistency SubsetEval Feature Selection Technique has higher accuracy than other Feature Selection Techniques. From Tables 7 and 8 it is observed that Consistency Subset Eval Feature Selection Technique is common and best method with higher accuracy for both the Tumor Datasets.

Hence, we conclude that Consistency Subset Eval Feature Selection Technique is best among fourteen Feature Selection Techniques for Tumor Datasets.

The performance of the Hybrid Method (C4.5 + MultiBoostAB) with best feature selection i.e., Consistency Subset Eval is compared with and without Feature Selection i.e., from our previous study [10] (Table 9).

There is slight improvement in the accuracy of the Hybrid Method with Feature Selection on PT dataset. Whereas nearly 6% improvement in the accuracy of the Hybrid Method with Feature Selection on CT dataset.

**Table 6** Other feature selection methods + MultiBoostAB + C4.5

Feature selection methods	Search technique	PT			CT				
		Reduced No. of attributes	Accuracy (%)	Time (s)	Tree (size)	Reduced No. of attributes	Accuracy (%)	Time (s)	Tree (size)
Chi-squared attribute Eval	<b>Ranker</b>	<b>17</b>	<b>43.66</b>	<b>0.13</b>	<b>85</b>	2000	82.26	3.97	5
Filtered attribute Eval	Ranker	17	43.36	0.13	85	2000	82.26	3.99	5
Info gain attribute Eval	Ranker	17	43.36	0.11	85	2000	82.26	3.97	5
Gain ratio attribute Eval	<b>Ranker</b>	<b>17</b>	<b>43.66</b>	<b>0.11</b>	<b>85</b>	<b>2000</b>	<b>83.87</b>	<b>3.97</b>	<b>5</b>
ReliefF attribute Eval	Ranker	17	43.36	0.31	85	2000	80.65	4.61	5
Principal components attribute Eval	Ranker	18	37.76	0.84	95	-	-	-	-
Symmetric Uncert Attribute Eval	<b>Ranker</b>	<b>17</b>	<b>43.66</b>	<b>0.11</b>	<b>85</b>	<b>2000</b>	<b>83.87</b>	<b>4.01</b>	<b>5</b>
OneR attribute setEval	FCBF	17	43.07	0.16	85	2000	82.26	6.91	5
Latent symantec analysis	Ranker	2	24.78	0.09	85	-	-	-	-

Bold indicates The best feature selection method with corresponding search technique



**Table 7** With all feature selection techniques for PT dataset + MultiBoostAB + C4.5

Feature selection techniques	Search technique	PT			
		Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size
Cfs subset Eval	Rank	11	41.89	0.11	81
Classifier subset Eval	Random	8	37.46	31.13	19
<b>Consistency subset Eval</b>	<b>Scatter</b>	<b>17</b>	<b>43.66</b>	<b>0.67</b>	<b>85</b>
Filtered subset Eval	Genetic	12	41.89	0.09	91
<b>Chi-squared attribute Eval</b>	<b>Ranker</b>	<b>17</b>	<b>43.66</b>	<b>0.13</b>	<b>85</b>
<b>Gain ratio attribute Eval</b>	<b>Ranker</b>	<b>17</b>	<b>43.66</b>	<b>0.11</b>	<b>85</b>
<b>Symmetric Uncert attribute Eval</b>	<b>Ranker</b>	<b>17</b>	<b>43.66</b>	<b>0.11</b>	<b>85</b>

Bold indicates The best feature selection method with corresponding search technique

**Table 8** With all feature selection techniques for CT dataset + MultiBoostAB + C4.5

Feature selection techniques	Search technique	CT			
		Reduced No. of attributes	Accuracy (%)	Time (s)	Tree size
Cfs subset Eval	Rank	47	80.64	43.64	5
Classifier subset Eval	Rank	1	83.87	6.64	–
<b>Consistency subset Eval</b>	<b>Rank</b>	<b>15</b>	<b>88.71</b>	<b>5.23</b>	<b>7</b>
Filtered subset Eval	Subset size forward selection	19	82.26	1.03	7
Gain ratio attribute Eval	Ranker	2000	83.87	3.97	5
Symmetric Uncert attribute Eval	Ranker	2000	83.87	4.01	5

Bold indicates The best feature selection method with corresponding search technique

**Table 9** Accuracy (%) of C4.5 with MultiBoostAB (Hybrid method) with and without feature selection technique

Data sets	Hybrid method (C4.5 + MultiBoostAB)	
	Without feature selection	With feature selection (consistency subset Eval)
Primary tumor (PT)	43.07	<b>43.66</b>
<b>Colon tumor (CT)</b>	82.26	<b>88.71</b>

Bold indicates The best feature selection method with corresponding search technique

## 5 Conclusion

Applying Feature Selection improves the standard of the information by removing immaterial attributes. In this study, various Feature Selection Techniques are considered for Tumor Datasets, various experiments are conducted. The performance of the Hybrid Method (C4.5 + MultiBoostAB) with various Feature Selection Techniques are compared in terms of accuracy, time to build a model and size of the tree on Tumor Datasets are observed. Experimental results show that Feature Selection technique enhances the accuracy of classification. From the results it is clear that, Consistency Subset Eval is the Best Feature Selection Method is identified for Tumor Datasets.

## References

1. J. Han, M. Kamber, *Data Mining; Concepts and Techniques* (Morgan Kaufmann Publishers, 2001)
2. M. Venkatadri. C. Lokanatha Reddy, A review on data mining from past to the future. *Int. J. Comput. Appl.* **15**(7), 19–22 (2011). ISSN: 0975-8887
3. Z.H. Zhou, Y. Jiang, Medical diagnosis with C4.5 rule proceeded by artificial neural network ensemble. *IEEE Trans. Inf. Technol. Biomed.* **7**(1), 37–42 (2003)
4. M. Lundin, J. Lundin, H.B. Burke, S. Toikkanen, L. Pylkkanen, H. Joensuu, Artificial neural networks applied to survival prediction in breastcancer. *Oncology* **57**, 281–286 (1999)
5. D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* **34**(2), 113–27 (2005)
6. M. Venkatadri, C. Lokanatha Reddy, A comparative study on decision tree classification algorithms in data mining. *IJCAETS* **2**(2), 24–29 (2010). ISSN: 0974-3596
7. R. Polikar, Ensemble based systems in decision making. *IEEE Circ. Syst. Mag.* **6**(3), 21–45 (2006)
8. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning application in cancer prognosis and prediction. *Comput. Struct. Biotech. J.* **13**, 8–17 (2015)
9. G. Sujatha, K. Usha Rani, Advanced ensemble technique on decision tree classifiers—an experimental study. *Special Issue Comput. Sci., Math. Biol. IJCSME-SCSMB* 264–268 (2016)
10. G. Sujatha, K. Usha Rani, Ensemble decision tree classifier performance with varying committee sizes. *Int. J. Comput. Eng. Technol.* **9**(1), 96–101 (2018)
11. W. Awada, R. Wald, A. Naplolitano, A review of the stability of feature selection techniques for bio informatics data, in *IEEE International Conference on Information Reuse and Integration* (IEEE, 2012), pp. 356–363

12. B. Pes, Feature selection for high dimensional data: the issue of stability, in *26th IEEE Conference, WETICE-2017* (2017), pp. 170–175
13. X.Y. Chen, B. Liu, Z.F. Zhang, X. Xia, The analysis of GCFS algorithm in medical data processing and mining. *Am. J. Softw. Eng. Appl.* **3**(6), 68–73 (2014). <https://doi.org/10.11648/j.ajsea.20140306.11>
14. C. Bala Kumar, A data mining approach on various classifiers in email spam filtering. *IJRASET* **3**(1), 8–14 (2015)
15. J.D. Novakovic, Support vector machine as feature selection method in classifier ensembles. *I.J. Mod. Educ. Comput. Sci.* **4**, 1–8 (2014)
16. B.C. Rokach, O. Maimon, Feature selection by combining multiple methods. *Adv. Web Intell. Data Min.* 295–304 (2019)
17. J.T. de Souza, N. Japkowicz, S. Matwin, Stochfs: a framework for combining feature selection outcomes through a stochastic process, in *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases* (2005), pp. 667–674
18. K. Polat, S. Sahan, H. Kodaz, S. Günes, A new classification method for breast cancer diagnosis: feature selection artificial immune recognition system (FS-AIRS), in *Proceedings of ICNC (2)'* (2005), pp. 830–838
19. C. Deisy, B. Subbulakshmi, S. Baskar, N. Ramaraj, Efficient dimensional reduction approach for feature selection, in *Conference on Computational Intelligence and Multimedia Applications* (2007)
20. M.A. Hall, L.A. Smith, Feature subset selection: a correlation based filter approach, in *1997 International Conference on Neural Information Processing and Intelligent Information Systems* (1997), pp. 855–858
21. S. Aruna, S.P. Rajagopalan, A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *Int. J. Comput. Appl.* **31**(8), 14–20 (2011). ISSN: 0975-8887
22. D. Lavanya, K. Usha Rani, Analysis of feature selection with classification: breast cancer datasets. *Indian J. Comput. Sci. Eng. (IJCSE)* (2011)
23. V. Kumar, Sonajharia, Feature selection: a literature review. *Soft Comput. Rev.* **4**(3) (2014)
24. M. Hall, Correlation-based feature selection for machine learning, Tech. Rep., Doctoral Dissertation, University of Waikato, Department of Computer Science, View at Google Scholar (1999)
25. UCI Irvine Machine Learning Repository, [www.ics.uci.edu/~mllearn/MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html)
26. Bioinformatics Group Seville, <http://www.upo.es/eps/big5/datasets.html>

# DNA Based Quick Response (QR) Code for Screening of Potential Parents for Evolving New Silkworm Races of High Productivity



K. Haripriya, D. M. Mamatha, S. Jyothi, and S. Vimala

**Abstract** India has a race privilege of having all varieties of natural Skills in the world. In such a demand, evolving new silkworm breeds specific to Resistance, Fecundity, Thermo tolerance and Silk productivity are always on demand. The conventional methods of silkworm breed Development is obsolete and the gene based scientific strategies are always on demand in screening potential parents for the cause of evolving new, sustainable and promising breeds. Genetic similarities of silkworm though collected from different research centres and Silkworm Germplasm bank cause a major concerns for this process. In view of this, molecular diversity is considered using DNA Barcoding through Co-I gene sequencing, DNA based QR code development and phylogenetic analysis. For this study Twenty nine silkworm races were analyzed. While DNA Barcoding is an effective molecular tool for Silkworm races and parental breeds identification, notwithstanding the advantages in DNA Barcoding, more specific DNA based QR codes are developed for the genetic identification and screening of silkworm races and parental breeds for the genetic improvement and promising breed development of silkworms.

**Keywords** Silkworm races · DNA barcoding · DNA based QR code · Python language

---

K. Haripriya (✉) · D. M. Mamatha · S. Vimala  
Department of Biosciences and Sericulture, Sri Padmavati Women's University, SPMVV,  
Tirupati, India  
e-mail: [Kodurharipriya@gmail.com](mailto:Kodurharipriya@gmail.com)

S. Vimala  
e-mail: [kodurharipriya@gmail.com](mailto:kodurharipriya@gmail.com)

S. Jyothi  
Department of Computer Science, Sri Padmavati Women's University, SPMVV, Tirupati, India

## 1 Introduction

The Quick Response code is a 2D barcode, which enables to encode more information than one dimensional barcode, prepared as a type of matrix barcode, which was first designed for the automotive industry by Denso Wave in Japan 1994 [1, 2]. It consists of Black and White Squares. These square patterns carry packets of information. It is based on 2 regions. They are Function patterns region and Encoding region. The function pattern that includes the finder pattern, timing pattern and alignment pattern deciphers the arrangement scheme where as the encoding region with the encoding data [3]. The appropriate orientation of code is detected by three squares displayed at the three corners of QR code symbol which are known as finder patterns. Decoder software enables to find the sides of pattern, Where as Alignment pattern used in correction of picture deformity.

The rest of the region is the encoded region where data code words and error correcting code words are stored. The quiet zone is the spacing provided to differentiate between QR code and its surrounding [3]. For the scanning program It is very important. Further, QR codes are fast readable codes and it can store information in the form of Uniform Resource locator (URL), plain text and numeric and image data information and etc. QR codes can efficiently provide such links for connecting collections, photographs, map; ecosystem notes citations and Gen Bank—NCBI sequences. QR codes have profuse advantages over DNA Barcodes.

In the present study, we have made a successful attempt to overcome the limitations and applications for reading the DNA Barcoding because of the difficulty in information retrieval through direct scanning of DNA sequences. The DNA sequence in plain format is very long string of character which is also not feasible for data input. We attempt to eliminate the limitation by encoding the DNA sequence into a more compact form. Quick Response code (QR) has been recognized as the best of the available barcode types for representing and is leveraged here to read the DNA sequences and to retrieve the genetic information and its taxonomic information too. In this study, a DNA based QR Codes were developed to widen the applications of DNA Barcoding technologies [4].

The QR code, similar to barcode, is an example of an information matrix. However a significant difference in the two is that while a DNA barcode only holds information in clear vertical bars, in one direction only, while a QR can hold information in both vertical and horizontal as well. This is why QR codes are referred to as two-dimensional, because they carry information both ways ie., vertically and horizontally. Another direct advantage to this is the great potential to carry pockets of information in a smaller space [2]. Compared to a DNA barcode, it has great scope of applications for quick identification.

## ***1.1 The DNA Barcode***

All organisms including all flora and fauna contain genomic DNA within the cell which consists nucleotide bases adenine, cytosine, guanine, thymine (A, T, G, and C) which are arranged in very specific sequences to encode the functional or structural proteins. DNA Barcode refers to a short section of DNA from a marker gene of the 'Mitochondrial genome'. Thus developed DNA barcode can be used to identify different species, races and pure breeds. Unlike fauna, plants have different gene markers to identify the different organism groups for DNA barcoding. The DNA sequences successfully used in DNA barcoding in animals is the 5' end of the mitochondrial gene Cytochrome oxidase 1 (COI). This sequence is recognized as universal barcode of the entire animal kingdom and it is used to authenticate and trace animal species races and pure breeds. The conventional means of generating DNA sequence data to obtain a barcode for a species or a specimen is through Polymerase Chain Reaction (PCR) amplification using species specific primers and sequencing of DNA barcode sequences through Sanger sequencing method from genomic DNA extracted from individual specimens [5].

## ***1.2 Sequence Alignment***

The DNA sequences of the specimens can be compared with the reference sequences to identify the species. BlastN Search of the NCBI—BLAST database with DNA sequence on sequence comparison algorithms. To identify regions of similarity between the DNA sequences, two sequences are aligned in an optimal arrangement. The optimal alignment of two sequences is chosen from the maximum score of matching pairs, mismatching pairs, and penalty score of the gaps. The current algorithm includes hierarchical clustering, similarity methods, combines clustering and diagnostic methods.

# **2 Materials and Methods**

## ***2.1 Generation of DNA Barcodes***

Following sequence alignment, from the tabulated DNA sequence files, Trace file and Taxonomy files are prepared and they are submitted to Barcode of Life Data system (BOLD) Database. After thorough verification and validations the accession numbers are specified for every DNA sequence. Then finally the respective DNA barcodes are generated by the BOLD systems [5].

## 2.2 *Generation of DNA Based QR Codes*

### **Anaconda-Python**

The Anaconda (Python distribution) is a standard platform for python programming language for scientific computing like Data sciences, data processing, predictive analytics etc. The Anaconda software is used by over 12 million users and includes more than 1400 popular data-science packages. It is suitable for windows, Linux and Mac OS. Anaconda distribution comes with more than 1400 packages as well as the conda package and virtual environment manager, called “Anaconda Navigator”.

Anaconda Navigator is a desktop graphical user interface (GUI) included in anaconda distribution that allows users to launch applications and environments and channels without using command-line commands. For DNA based QR codes development “JUPYTER NOTEBOOK” application has been used. This application is available by default in Navigator [6].

Jupyter is a browser-based interpreter that allows you to interactively work with Python. You can think of jupyter as a digital notebook that gives you an ability to execute commands, take notes and draw charts. It’s primarily used by Data Scientists.

From the Launcher tab, open the Python 3 kernel in the Notebook area. A new Jupyter notebook file with an empty code cell opens in a separate tab. Enter your python program in the code cell. To run the program and add a new code cell below the program, select the cell in the notebook from the toolbar.

## 2.3 *Data Set of DNA Sequences*

The COI gene sequence of the selected specimens were retrieved from BOLD Database. The DNA sequence were Queried from the BOLD Database by keywords “barcode”. Some of the DNA sequences were extracted from the complete genome as CDS for gene COX I. The DNA barcode sequences of all silkworm Bombyx breeds were chosen as reference barcodes. The DNA sequence with BOLD index numbers: AAB3839 were chosen as test set (Different silkworm breed samples) [4].

## 2.4 *DNA Sequences into DNA QR Code Encoding*

The QR code has the best compression efficiency in encoding DNA barcode sequences among the other 2D codes (Matrix) [7]. The open source QR Code Library in Python kernel version (Anaconda) was adapted for developing program in Jupyter notebook to encode the DNA sequences [8]. To generate QR code, a new python jupyter notebook in kernel version is created at first in order to write QR code collecting commands for running the programme. At this stage it is called encode ( ).

In the encode method with four settings; the text to the DNA sequence (encode), QR code (barcode type) and the desired width and height (in pixels) of the image are produced. The applications general commercial QR codes are endless. They are used to identify and classify magazines, advertisements, product wrappings, T-shirts, passports, business cards etc [1].



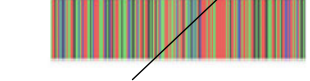



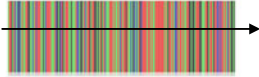
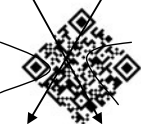
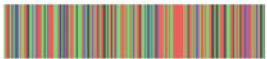

In the present study the silkworm races and breeds obtained from the germplasm bank of Andhra Pradesh State Sericulture Research Institute (APSSRDI) are considered. The genomic DNA of the marker gene sequence (Co Igene) is taken as input sequence. In addition to the DNA sequence, its taxonomic data and breed name are also taken as inputs. Applying the above programme the DNA based QR code is developed for every respective Silkworm races and breeds.

## 2.5 Comparison Between DNA Barcode and DNA Based QR Code

S. No.	DNA barcode	DNA based QR code
1.	A DNA Barcode only holds information in the vertical direction	QR code holds information in both horizontal and vertical directions
2.	A DNA Barcode can only hold as much information that could be limited to the number of specific DNA bases and position and the number of DNA bases	The capacity of QR code is hundreds of times higher in storing and linking the information than a DNA barcode [3]
3.	DNA barcode takes up Alphabetical information only	While QR code takes up information in Alphanumeric
4.	Taxonomic information and meta data cannot be included	Organism's taxonomic information can be included consisting of entire binomial nomenclature, including species type and breed name and common name too
5.	Image data and meta data cannot be contained in a DNA barcode	While QR code can be generated not only with Alphanumeric but also image data and meta data too
6.	DNA barcode scanners are not available and need to be developed specifically	While the QR code can be scanned with any QR code scanner publicly available and hence can reach out to the common man

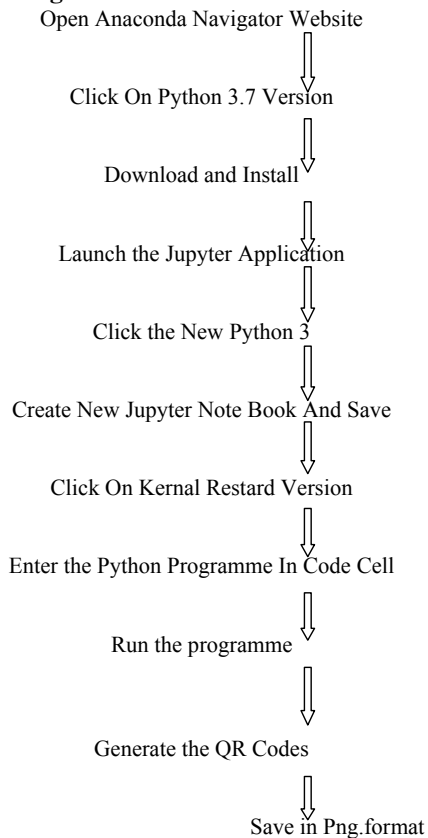
See Fig. 1.



Features of the code	DNA barcode	DNA based QR code
High capacity	650-680 bases Alpha 	Up to 7089 numeric integers 
Durability against soil and damage	Reading is impossible 	Reading is possible ( upto 30% damaged) 
Reduced space	10 digits numeric (approx 50 mm X 20 mm) 	40 digits Numeric (approx 5mm X 5mm) 
360° reading	Horizontal reading 	Supports 360° reading 
Language supported	Alpha/sequence based ATGCGACTACTCGCATGCTTG 	Numeric alphanumeric ,kanji, kana #%\$@& 123AGB x,y,z €£¥ 

**Fig. 1** Features of DNA barcode compared to DNA based QR codes

### Program Flow Chart



## 3 Results and Discussion

The DNA based QR code has the capacity to hold diverse types of data: Numeric and Alphabetic characters, Kanji, Hiragana, Katakana, symbols, binary and control codes. The QR Code has the capacity to encode 7089 numeric version and 4296 alpha numeric characters that is hundreds of times more in comparison with the present barcode. The QR code, though one tenth smaller than a regular DNA barcode, holds larger information, with high speed scanning capacity, Omni-directional readability and having “position detection patterns” that circumvents the negative effects of background interference. Error correction is possible. It has the ability to secure the information against 30% damage. It provides security upto four levels viz H (30%), Q (25%), M (15%) and L (7%) security limits. Further it possesses an additional feature of Compartmentalization that divides multiple data areas resulting in tiny printouts. It can readable on any Android, iOS, window versions.

### ***3.1 DNA QR Code of Silkworm Breeds***

DNA based QR codes of the samples look darker than the common QR codes due to pixel density.

In the present study DNA based QR code of 9 different silkworm races and breed from the APSSRDI were taken. As seen in Fig. 2. More data we put into QR code, the more rows and columns of modules (the little black squares) will be introduced. Therefore the minimum width of printed QR code image depends on the size of individual module when viewed by the camera (from a distance at some resolution). We found that the minimum width of a printed DNA QR code is 2.8 cm. Smaller than that minimum width, the scanner cannot read the DNA QR code correctly.

### ***3.2 Scanning the DNA QR Code***

**DNA based** QR code must be scanned with a Smartphone (matrix can be read quickly by a cell phone) using QR code scanner app. For some applications it requires internet connection. When the application has started, it shows blank rectangle in the middle of the device screen and captioned buttons place at the bottom. By touching the scan button, the screen switched to the camera interface. User can place the DNA QR code inside the viewfinder. The scanner captured the DNA QR code and decoding back into DNA sequences including its taxonomical data and its race/breed name. If the scanner failed to decode the DNA QR code, the rectangle contains some random numbers only. In this case, the user should repeat the scanning again for deciphering the coded information.

## **4 Conclusion**

This study enabled a successful attempt to overcome the limitations and applications of deciphering the DNA Barcoding because of the difficulty in information retrieval through direct scanning of DNA sequences. The DNA sequence in plain format is very long string of characters represented by the nucleotide bases which is also not feasible for data input. We attempt to eliminate the limitation by encoding the DNA sequence into a more compact form. DNA based Quick Response code (QR) has been recognized as the best of the available barcode types for representing and is leveraged here to read the DNA sequences and to retrieve the genetic information and its taxonomic information too. In this study, a DNA based QR Codes were developed to widen the applications of DNA Barcoding technologies. Another direct advantage to this is great potential to carry pockets of information in a smaller space. Compared to a DNA barcode. This technology enables even the common man to technically identify the various organisms at its best correctness right from the

S.No.	Acc./code No & Type	DNA Sequence of marker gene: Mitochondrial CO-I Primers : C_LepFolF & C_LepFolR	DNA based QR CODE
1	<p><b>APS45 &amp; Chinese</b></p>  <p>Larval markings: Plain Cocoon shape : oval Cocoon colour : white Cocoon size :Medium</p>	<p>ATTGATCACGCATAATTGGAACATCTTTAAG ACTTTTAAATCGAGCTGATTTAAGAAATCCA GGATCATTGATGGAGATGATCAAATTTATA ATACCTATTGTAACAGCACATGCTTTTATTATA ATTTTTTTTATAGTTATACCTATTATAATGG AGGATTTGGAATGATTAGTTCCTCTTATAC TAGGAGCACCAGATATAGCATTCCCAGCAAT AAATAATATAAGATTTGACTCCTACCCCTC CCCTTATATTATAATTCAAGAAGAATTGTA GAAATGGTGCAGGAACAGGATGAACAGTTT ACCCCCACTTTCACTAATATCGCACATAGA GGAAGATCCGTAGATCTTGCTATTTTTTCACT ACATTTAGCAGGTATTTCACTAATTAAGGA GCAATTAATTTTATTACAACAATAATTAATAT ACGATTAATAATATATCATTGATCAAITTAC CCTTATTTGATGAGCTGAGGGATTACAGC ATTTTTATTATTATCACTACTGTTTTAGC TGGAGCTATTACAATATTATTAACAGATCGA AACTTAAATACATCTTTTGAACCCGGGG GAGAA</p> <p><b>DNA Barcode</b></p> 	
2	<p><b>APS12 &amp; Japanese</b></p>  <p>Larval markings: Plain Cocoon shape :Dumbbell Cocoon colour : white Cocoon size :Medium</p>	<p>AATTGGAACATCTTTAAGACTTTTAATTCGAG CTGAATTAGGAAATCCAGGATCAITTAATGG AGATGATCAAATTTATAATACATTTGTAACA CGCATGCTTTTATTATAATTTTTTTTATAGTT ATACCTATTATAAATGGAGGATTTGGAATTT GATTAGTTCCTTATACTAGGAGCACCAGA TATAGCATTCCCACGAATAAATAATATAAGA TTTTGACTCCTACCCCTCCTTATATTATT AATTTCAAGAAGAATTGAGAAATGGTGCA GGAACAGGATGAACAGTTTACCCCCACTTT CATCTAATATCGCACATAGAGGAAGATCCGT AGATCTTGCTATTTTTTCACTACATTTAGCAG GTATTTCACTAATTTAGGAGCAATTAATTTT ATTACAACAATAATTAATATACGATTAATAA ATATATCAITTTGATTCAATACCCTTATTTGT ATGAGCTGAGGGATTACAGCATTITTTATAT TATTACTACTCTGTTTTAGCTGGAGCTAIT ACAATATTATTAACAGATCGAACTTAATA CATCAITTTTGTATCCTGCTGGAGGAGGAGA CCCAATTTTATATCAACATTTTATT</p> <p><b>DNA Barcode</b></p> 	 <p>APS12 &amp; Japanese</p>
3	<p><b>APDR105 &amp; Chinese</b></p>  <p>Larval markings: Plain Cocoon shape : oval Cocoon colour : white Cocoon size :Medium</p>	<p>TAAGACTTTTAATTCGAGCTGAATTAGGAAA TCAGAGATCATTAAATGGAGATGATCAAAT TATAATACTATTGTAACAGCACATGCTTTTAT TATAATTTTTTTTATAGTTATACCTATTATA TTGGAGGATTTGGAATGATTAGTTCCTCTT ATACTAGGAGCACCAGATATAGCATTCCCAC GAATAAATAATATAAGATTTTGACTCCTACC CCCTCCCTTATATTATAATTTCAAGAAGAA TTGAGAAAATGGTGCAGGAACAGGATGAAC AGTTTACCCCTTCACTCAATATCGCAC ATAGAGGAAGATCCGTAGATCTTGCTATTTT TCACTACATTTAGCAGGTATTTCAATCAITAT AGGAGCAATTAATTTTATTACAACAATAAAT AATATACGATTAATAATATATCATTGATC AATTACCCTTATTTGATGAGCTGAGGGATT ACAGCATTITTTATTATTATCACTACTCTGT TTTAGCTGGAGCTATTACAATATTATTAACAG ATCGAAACTTAAATACATCTTTTGTATCCT GCTGGAGGAGGAGA</p> <p><b>DNA Barcode</b></p> 	 <p>APDR105 &amp; Chinese</p>

Fig. 2 Total input information (external morphology, classification, CoI gene sequence, DNA barcode, race/breed name) for the development of DNA based QR code









<p>4</p> <p><b>APS71 &amp; Chinese</b></p>  <p>Larval markings: Plain Cocoon shape : oval Cocoon colour : white Cocoon size :Medium</p>	 <p>AACATTATATTTTATTTTGGTATTGATCAG GAATAAATGGAAACATCTTAAAGACTTTTAAAT CGAGCTGAATTAGGAAATCCAGGATCATTAA TTGGAGATGATCAAATTTATAACTATTGTA ACAGCATGCTTTTATTATAAATTTTTTTTAT AGTTATACCTATTATAAATGGAGGATTGGA AATTGATTAGTTCCTTACTAGGAGCACC AGATATAGCATTCCCAGAAATAATAATA AGATTTGACTCCTACCCCTCCCTTATATT ATTAATTTCAAGAAGAAATGTAGAAAATGGT CAGGAACAGGATGAACAGTTTACCCCCAC TTTCATCAATATCGCACATAGAGGAAGATC CGTAGACTTGCTATTTTTCATACATTAG CAGGTATTCATCAATATAGGAGCAATTA TTTTATTACAACAATAATTAATACGATTA ATAATATACATTTGATCAATTACCTTATTT GTATGAGCTGAGGGATTACAGCATTTTTATT ATTATTACTACCTGTTTATAGCTGGAGCTA TTACAATATTATTAACAGATCGAAACTTAA TACATCATTTTTATGCTGCTGGAGGAGGA GACCAATTTTATCAACATTTATTT</p> <p>DNA Barcode</p> 	 <p>APS71 &amp; Chinese</p>
<p>5</p> <p><b>APS72 &amp; Japanese</b></p>  <p>Larval markings: Plain Cocoon shape :Dumbbell Cocoon colour : white Cocoon size :Medium</p>	<p>AACATCTTTAAGACTTTTAATTCGAGCTGAAT TAGGAAATCCAGGATCAATTAATGGAGATGA TCAAATTTATAATACTATTGAACAGCACAT GCTTTTATTATAAATTTTTTTATAGTTATACCT ATTATAATGGAGGATTGGAAATTTGATTAG TTCCTCTTACTAGGAGCACCAGATATAGC ATTCCCAGAAATAATAATAAGATTTTGA CTCCTACCCCTCCCTTATATTTAATTTTC AAGAAGAATTGTAGAAAATGGTGCAGAAC AGGATGAACAGTTTACCCCCACTTTCATCTA ATATCGCACATAGAGGAAGATCCGTAGATCT TGCTATTTTTCACTACATTTAGCAGGTATTT CATCAATTATAGAGGCAATTAATTTTATTACA ACAATAATTAATATACGATTAATAATATAT CAITGATCAATACCTTATTTGTATGAGCT GTAGGGATTACAGCATTTTTATTATTATTATC ACTACCTGTTTATAGCTGGAGCTATTACAAT TATTAACAGATCGAAACTTAAATACATCATT TTTTGATCCTGCTGGAGGAGGAGACCAATT TTATATCAACATTTATTT</p> <p>DNA Barcode</p> 	 <p>APS72 &amp; Japanese</p>
<p>6</p> <p><b>APS33 &amp; Chinese</b></p>  <p>Larval markings: Plain Cocoon shape : oval Cocoon colour : white Cocoon size :Medium</p>	<p>ATTATGTGTCACACACTCACAGATATTCGGA CCATATATATTATTGTTGGTTCATCCAGGA CTAATGGGAACATCTTAAAGCTTTTTAATTCG ACGCTGATTTAAGGAATCCAGGATCAITTAATT GGAGATGATCAAATTTATAACTATTGGAA CAGCGACGCTTTTATTATAAATTTTTTTATA GTTATCCCTATTATAATTGGAGGATTGGAA ATTGATTAGTTCCTCTTACTAGGAGCACC GATATAGCATTCCCAGAAATAATAATA GATTTGACTCCTACCCCTCCCTTATATTA TTAATTTCAAGAAGAATTGTAGAAAATGGT CAGGAACAGGATGAACAGTTTACCCCCACT TTCATCTAATATCGCACATAGAGGAAGATCC GTAGACTTGCTATTTTTCATACATTAGC AGGTATTCATCAATATAGGAGCAATTAAT TTTTATTACAACAAGAAGTAAATACGATTA ATAATATATCAITTTGATCAATTACCTTATTT GTATGAGCTGAGGGATTACAGCATTTTTATT ATTATTACTACCTGTTTATAGCTGGAGCTA TTACAATATTATTAACAGATCGAAACTTAA</p>	

Fig. 2 (continued)



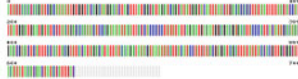







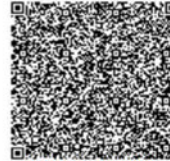

		<p>TACATCAITTTTTGATC</p>  <p>DNA Barcode</p>	
7	<p>APSHT02 &amp; Chinese</p>  <p>Larval markings: Plain                  Cocoon shape : oval                  Cocoon colour : white                  Cocoon size :Medium</p>	<p>AGAGTGGACATCTTTAAGACTTTTAATTGCGA                  GCTGGAATTAGGAAATCCAGGATCATTAAATG                  GAGATGATCAAATTTATAACTATTGTAAC                  AGCACATGCTTTTATTATAATTTTTTTATAG                  TTATACCTATTATAAATGGAGGATTTGGAAAT                  TGATTAGTTCCTTTATACTAGGAGCACCAG                  ATATAGCATTCCCACGAATAAATAATATAAG                  ATTTTGACTCCTACCCCCCTCCTTATAITAT                  TAATTTCAAGAAGAATTGTAGAAAATGGTGC                  AGGAAACAGGATGAACAGTTTACCCCCACTT                  TCATCTAATATCGCACATAGAGGAAGATCCG                  TAGATCTTGCTATTTTTCACTACATTTAGCA                  GGTATTTCAATTAATAGGAGCAITTAATTT                  TATTACAACAATAAATAATACGATTAAT                  AATATATCATTGTATCAATTACCCCTATTGTT                  ATGAGCTGAGGATTACAGCATTTTTTATAT                  TATTACTACTGTTTATAGCTGGAGCTATT                  ACAATATTATTAACAGATCGAACTTAATA                  CATCAITTTTTGATCCTGCTGGAGGAGGAGA                  CCCAATTTATAATCAACATTTATTTGATTTT                  TG</p> <p>DNA Barcode</p> 	
8.	<p>APSHT02 &amp; Japanese</p>  <p>Larval markings: Plain                  Cocoon shape :Dumbbell                  Cocoon colour :white                  Cocoon size :Medium</p>	<p>GGACATCTTTAAGACTTTTAATTCGAGCTGA                  ATTAGGAAATCCAGGATCATTAAATGGAGAT                  GATCAAATTTATAACTATTGTAACAGCGC                  ATGCTTTTATTATAATTTTTTTATAGTTATAC                  CTATTATAATTGGAGGATTTGGAAATGATT                  AGTTCCTTATACTAGGAGCACCAGATATA                  GCATTCCCAGGAATAAATAATATAAGATTTT                  GACTCCTACCCCCCTCCTTATAITTAATTT                  TCAAGAAGAATTGTAGAAAATGGTGCAGGA                  ACAGGATGAACAGTTTACCCCCACTTTTAT                  CTAATATCGCACATAGAGGAAGATCCGTAGA                  TCTTGCTATTTTTCACTACATTTAGCAGGTA                  TTTCAATTAATAGGAGCAATTAATTTTTATT                  ACAACAATAATGAATATACGATTAATAATA                  TATCATTGTATCAATTACCCCTATTGTATGA                  GCTGTAGGATTACAGCATTTTTATTATTATT                  ATCACTACCTGTTTTAGCTGGAGCTATTACAA                  TATTCTTATC</p> <p>DNA Barcode</p> 	
9	<p>CTIPP &amp; Chulthai</p>  <p>Larval markings: Plain                  Cocoon shape :Dumbbell</p>	<p>TTTGTCCGAAAATTTGGGACATCTTTAAGA                  CTTTTAATTCGAGCTGAATTAGGAAATCCAG                  GATCAATTAATGGAGATGATCAAATTTATAA                  TACTATTGTAACAGCACATGCTTTTATTATAA                  TTTTTTTATAGTTATACCTATTATAATTGGA                  GGATTTGAAAATGATTAGTTCCTTTATACT                  AGGAGCACCAGATATAGCATTCCCACGAATA                  AATAATAAGATTTGACTCCTACCCCCCTC                  CCTTATTATTAAATTTCAAGAAGAATTGTAG                  AAAATGGTGCAGGAAACAGGATGAACAGTTT                  CCCCCACTTTCATCTAATATCGCACATAGAG                  GAAGATCCGTAGATCTGCTATTTTTCACTA                  CATTAGCAGGATTTTCAATCAATTAAGGAG                  CAATTAATTTTATTACAACAATAAATAATA</p>	

Fig. 2 (continued)

	Cocoon colour : white Cocoon size :Medium	<pre>CGATTAATAATATATCATTTGATCAATTACC CTTATTGTATGAGCTGTAGGGATTACAGCAT TTTTATTATTATTACTACTGCTTTTAGCTG GAGCTATTACAATATTATTAACAGATCGAAA CTTAAATACATCATTTTTGTCCCTGCTGGAG GAGAGA</pre> <p>DNA Barcode</p> 	
--	--	---	--

**Fig. 2** (continued)

external morphological features to gene level identification. Thus this study helps to screen and identify the potential parents for evolving new silkworm races. This technology can obviously be applied for other organisms too.

## References

1. T. Nauli, *DNA QR Code Scanner for Identifying the Species Origin of Meat Products* (2015)
2. F.H. Yahya, H.A.R.L. Yussof, *Integration of Screen Cast Video Through QRcode: An Effective Learning material for m-Learning* (2018), pp. 1–13
3. K.H. Pandya, H.J. Galiyawala, *A Survey on QR Codes: in Context of Research and Application* (2014), pp. 258–262
4. S. Vimala, D.M. Mamatha, G.D. Khedkar, J. Raju, *DNA Barcoding Studies of Different Mulberry Silkworm (Bombyxmori) Breeds and their Phylogeny Based on Computational Tools* (2016), pp. 1–5
5. S. Kundu, *Molecular Phylogeny of South Indian of Prawn Species by DNA barcoding using COI gene as a Marker* (2016), pp. 56–59
6. <https://anaconda.org/conda-forge/qrcode>, <https://www.anaconda.com/distribution/>
7. M. Rajagopal, *Security Empowerment Using QR Code and Session Tracking for Cued Recall Based Textual Password Users* (2018), pp. 2325–2329
8. G. Durak, E. Ozkeskin, M. Ataizi, *QR codes in Education and Communication* (2016), pp. 42–58

# Identification of Neighbourhood Cities Based on Landuse Bigdata Using K-Means and K-NN Algorithm



S. VinilaKumari, P. Bhargavi, and S. Jyothi

**Abstract** In present days, several cloud computing platforms or web services such as Flip kart and Amazon, Google App Engine, blue cloud etc. provide a locally distributed and scalable data which is in uncountable form. But, these platforms do not regard geographical location data. However, the data is generated from the modern remote satellites with their geological topology. The so obtained geo-distributed database is able to process either a large scale data or a very simple type, scalable while being fault-tolerant and fast in answering a query. The processing of Big data includes the storing and analysing the uncountable amount of geographical data. The big data processing utilizes several programming models and frameworks such as Map Reduce, Hadoop, MongoDB, Pig etc. The present work concentrates on land use classification of various cities in India using geographical location data having latitude and longitude of every boundary. To perform this work, India map shape file with every state is used. The shape file is converted into longitude and latitude band information along with cities data. Nevertheless, the geo-graphical data is classified by applying the machine learning algorithms.

**Keywords** Big data · Geographical location data · K-NN algorithm · Clustering

---

S. VinilaKumari (✉) · P. Bhargavi · S. Jyothi  
Department of Computer Science, SPMVV, Tirupati, India  
e-mail: [vinisd2006@gmail.com](mailto:vinisd2006@gmail.com)

P. Bhargavi  
e-mail: [pbhargavi18@yahoo.co.in](mailto:pbhargavi18@yahoo.co.in)

S. Jyothi  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_10](https://doi.org/10.1007/978-3-030-46939-9_10)



## 1 Introduction

The land usage/classification [1] is done to understand the land usage problems for development purposes by using aerial photographs. USGS, United States Geological Survey devised a land cover/land use classification [2] system for use with the remote sensing data. USGS is a main resource of Geographic Information System (GIS) [3] data.

In the recent past the rampant use of location-based services, as well the usage of technologies that are dependent on location/activity, have led to the generation of massive amount of big data that is location-based. Even data comprising of tracking or sensing data (e.g., data from GPS trajectories fitted in the vehicles, and even geo-referenced mobile phone data), apart from the data seen on the social media like Twitter.

**Geospatial Technology** is an emerging field of study that comprises of Geographic Information System (**GIS**), Remote Sensing (**RS**) [4] and Global Positioning System (**GPS**). It uses systems that acquire and handle location-specific data about the earth and use it for analysis, modelling, simulations and visualization. It allows us to make informed decisions based on the importance and priority of resources most of which are limited in nature. It may also be used to create intelligent maps and models that are interactive and can be queried to get the desired results in a STEM (Science, Technology, Engineering, and Math) application or may be used to advocate social investigations and policy-based research. It may be used to reveal spatial patterns that are embedded in large volumes of data that may not be accessed collectively or mapped otherwise. Geospatial technology has become an essential part of everyday life. It's used to track everything from personal fitness to transportation to changes on the surface of the earth.

The rampant growth of such geospatial big data has created exceptional chances for analysis on urban systems and human environments. Meanwhile, it additionally challenged us with more avenues questions in hypothetical, technological, moral and social questions.

Machine learning algorithms (MLA) [5] proved as successful development methods for the last few years. The algorithms especially complement the support in search engines besides the speech recognition systems and robotics which use several data intensive systems and object detection and soon. Over the few decades, studies reveal that for securing information on land cover classification; there exists identification difficulty in habitual classification because of lack of spectral information. Therefore, to improve the accuracy of classification in different environments, many studies have integrated machine learning algorithms. The current paper is on how; the GIS data is collected from shape file as Big data then loaded into the MongoDB. The data is classified using MLA to locate the major cities in India and its boundary is identified using MLA.

## 2 Big Data and Geographic Information System

### 2.1 Big Data

Big data is defined as datasets containing various data styles to analyse the data that is tough to be handled and analysed with the active data processing systems. The features of big data are characterized into '3V', called, volume, variety and velocity [6].

Firstly, big data is used to deal massive datasets, typically greater than terabytes derived from GPS, (Global Positioning System) volumes of social media, and other resources. Terabyte stands for units of data equal to one million \* million ( $10^{12}$ ) bytes, or 1024 GB. The 'big data' on the whole imply vast size of datasets which is incredibly vast in contrast to the precedent datasets.

Secondly, big data hosts datasets that are varied in pictures, sounds, video streaming, maps and also text messages of social media. However, Big data besides targeting the structured datasets it also targets formless ones that were typically out of interest to people that deal with data. It is quite varied and is away from our mind's eye and integrates completely diverse sorts of datasets to get fresh sort of information. The systems of Big data utilize a computer cloud and even other platforms like Hadoop for information integration and amalgamation.

Big data's third characteristic is its extreme velocity in producing, spreading, and applying the results in the real time. Besides, Big data's process of analysing can be augmented with social network services such as Face book or Twitter [7]. These platforms consider the photos posted by individuals as datasets, which present live evidence of location, liking, and other individual information that is useful for analytics and business promotions. Besides, this piece of information will also be utilized for advertising and sales promotion by various businesses or even to propagate the policy measures by government.

Big data although in its narrow definition highlights the source of data, compiling, storage and other technological subjects, in its broad definition includes analysis and expression aspects. In other words, big data is defined as a massive tool containing datasets that are differently formatted hosting analytic methods to process social network services, as well as statistical synthesis along with revelation of picture. Major components of big data include resource, technology, and human capital [6]. Resource stands for high quality data acquisition. While the technology indicates its stand referring to storing of data, its managing, processing, analysis, and visualization, Human capital denotes the data scientists who possess expertise in mathematics, engineering, economics, statistics and psychology. Besides, these data scientists communicate with people are in making a disclosing the result in the form of a creative story and even effectively visualizing the big data contents.

## 2.2 *Big Data and Geographic Information System*

GIS is all a virtual world, comprising of points, polygon, line and graph. Ever since the establishment of GIS as a field, analysing of the datasets constantly remained a challenge. Nevertheless, analysing of massive data has always remained enduring problem in both conventional Information and Technology(IT) as well the Geo-Spatial domain. However, owing to the latest growth in infrastructure pertaining to both hardware and software, the dispensation of vast data sets made easy. Thus a big thrust and new direction was given to industries that were hampered by slow data dispensation abilities. GIS however, augmented the industry utilizing the chance. According to McKinsey report, development in Big Data [8] will set a new wave of innovation. This innovation would be felt all across the IT sector. Innovation in Big Data and GIS [9] will bring in a lot of new players into the market.

GIS possesses its own taxonomy pertaining to Big Data analysis. Enormous data sets are termed as Spatial Big Data (SBD). Big Data is conventionally defined by 3V's: Volume, Velocity and Variety [10]. On the other hand, Spatial domain is confronted with the problem of enlargement in size, diversity and update occurrence, which surpassed the competence of the normally used spatial calculating techniques, design, methods and software solutions.

The spatial data obtained is usually in Raster, Vector and Graph; SBD dominates in the said genres over the time and it hosts data types like satellite imagery, climate simulation, multiple and coordinated drone imaging system. Vector type data includes that of Uber, location specific twitter data, GPS tacking data etc. are present in SBD. Besides it also includes other data types like supply-chain network data, electric-grid data, graph data, road-network data, network data of drones.

However, SBD has its own brevities like lack of specialized systems, technical methods and algorithms to support every data type of SBD. Indeed, highly sophisticated tools and concepts of Big Data [10] like Map-Reduce technique, Hadoop software, Hive, HBase, Spark do not support spatial or Spatio-temporal data directly. However most of SBD are analysed as a non-spatial data or using a wrapper function which fails to bring down data dispensation time though.

Big data and GIS have many aspects in common owing to their similarities in parts of data analysing. Open source and even commercial software along with web based online GIS systems are available for processing GIS data as on date.

Firstly, GIS uses information containing the location or area, displaying it in the form of a picture or a map. Currently, satellite information is playing a pivotal role in the latest new technologies available. GIS data is often large-sized similar to Big data as it is primarily a location based information.

Secondly, GIS also collects survey information like street data, CCTV footage, or any other such location-based data. If by chance, the datasets failed to provide location-based information, the GIS programmers do a geo-coding procedure to find the location and convert the same into data sets of GIS. Besides, people's coordination

is also significant to acquire the GIS data; hence, the democratic GIS system happens to be a huge field of GIS. Swarming with programme mechanism is a big tool for acquiring the information in GIS.

Thirdly, GIS hosts either internet server, a geospatial data server, or a cloud server for storing the data. These servers can sometimes overlap one over the other but with a restriction that their own territories pertaining to sharing. The fundamental principle of geo-database for single-user and multi-user according to the ESRI's official website information is Geo-database system which is very essential to administer complex structured GIS datasets along with their features.

Fourthly, the online software pertaining to GIS desktop plays a major role in the remaining process as well as data processing (building), analysis, and visualization. In the GIS data processing (building), competent systems like ArcGIS Online, Google Maps JavaScript API included are Maps JavaScript API, Microsoft Bing Geocode Dataflow API, and US Census Geocoder. These play helping role for constructing geo-coding and mapping coordinates in the database.

Finally, GIS data analysis hosts quite a few functions with ArcGIS analysis toolbox. Similarly, it hosts even in softwares such as ArcGIS, QGIS, GRASS GIS, GeoDa, CartoDB, Mapbox, and the other desktop or online GIS systems.

### ***2.3 Big Data's Data Process and Analysis Techniques***

In Big data processing, more technologies are developed and are categorized into data processing concepts. In processing a lot of content, the content must be extracted and analysed from the collected information to serve the knowledge requirements of various business organizations, political parties and scientific research departments. The process is initiated with the retrieval of information, which can come from various sources like database, websites, documents or content management system. Hadoop [11] is responsible for storing massive amount of data.

Before preparing big information it must be recorded from different information creating sources. In the wake of chronicle, it must be filtered and optimized. Just the pertinent information ought to be recorded by a method of channels that dispose the futile data. To encourage this work, specific instruments are utilized, for example Extraction, Transformation, Loading (ETL) method is considered which combines data from multiple systems in which the data is actually loaded into the data state. The ETL flow is as shown in Table 1.

Figure 4 shows big data showcases parts within which the method has information supply, collection, storage, processing, with analysis and visualization. In the methodology described below, every step comprises of parts not similar to each other from the past info systems that typically reserved structured datasets.

Firstly, the source of information for big data's is generally from institutions or from the internal database of an organization or from external sources like Twitter

**Table 1** Stages in ETL approach

S. No.	Stages	Description of stages
1	Extraction	The first segment of ETL procedure is Extraction of the information from the source framework. However, accurately separating the information places the phase for the accomplishment of ensuring procedures. The majority of the undertakings are to consolidate information with a few distinctive source frameworks
2	Transformation	In this phase rules or principles are applied to the extracted information loads into the final target. If any information needs no change whatsoever, then that information is called as “immediate move” or “go through information”
3	Loading	The loading stage loads huge volume of data, loaded in a short period and should be optimized for better performance

or Face book, or pictures and any streaming videos from internet. Usually, urban geographic researches and projects utilize a large scale spatial database [12], called big data.

Secondly, big data utilizes a crawling-method along with a search-engine to get data from Internet in the assortment method. It even utilizes Internet of Things (IoT) based sensors to gather the data. This itself creates a large difference to big data from the past traditional data assortment methods.

### 3 Big Data as Another Visual Image Tool for GIS

Over the time, the trends in big data have radically hit the industry and it's not a big shock that big data in GIS has considerable repercussions on how we obtain and influence spatial information. On looking at the methods Industries utilize geographic information, it is quite obvious that usage has expanded rapidly over the time; however, in the past, only government agencies were the largest adopters of geospatial data. But now it is clear that widespread adoption to GIS is found in Industrial corridors. The convergence of GIS [13] with big data means that the potential applications of the two will become limitless. We wanted to look at why interest appears to have risen so dramatically and how different organizations are using big data together with GIS.

The visual image tool in Big data for GI Scan be drawn from many instances in Big data. Big data and GIS share some of the aspects in the visual image and demonstration technology. However, they have at least one exclusive aspect that cannot be mutual (see Fig. 1). The figure hosts three colours presenting: (A) presenting the exclusive area for GIS visualization, (C) presenting exclusive are of big data visualization, and (B) an overlapping area of the two technologies. In the figure (A) indicates that visualization shows supported location or map with geographic

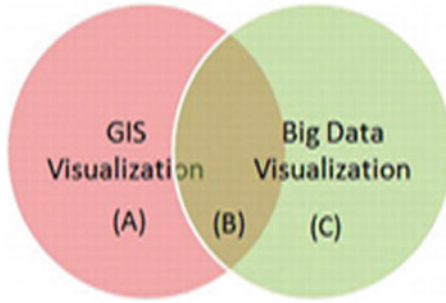


Fig. 1 Venn diagram of GIS and big data visualization



Fig. 2 GIS visualization example of US cities

coordinates. Meanwhile, the select area of big data visualization (C) shows an image demo without a location or a map, which does not denote spatial context.

Many big data visualizations present the results that belong to the exclusive area as they do not have any geographic qualities or variables. (A) in Fig. 2 is an example, while Fig. 3 is a fine instance of the area (C). In Fig. 2 pertaining to US cities, the large bubble implies higher city location in visualization mapping technology. Regardless the spatial context or geographic coordinates, this is a clean big data visual image area. However, it is found that the overlapping areas (B) are vast in data. Hence big data visualization technologies are utilized for storing and processing the data. In Fig. 4, presents common area (B) with the Chernoff face and the US map, in which the Chernoff face indicates multivariate big data image utilizing variables resembling human face with SAS or R or python programming. There are several substitute visual image examples offered for big data terms if rooted in maps or spatial context. Figure 4 is also a fine instance of area (B) as it presents the position without using a map.



Fig. 3 Example of big data picture of gender and ethnicity in tech companies with online tableau public [14]



Fig. 4 US states death penalty executions since 1976 [15]

### 4 Machine Learning Algorithm

A revision of algorithms and statistical models to do a particular task without any clear instructions but by just depending on some prototype and inference is known as Machine Learning. Indeed, a mathematical model of sample data or “training data” is constructed using Machine Learning Algorithms to do forecast and even make conclusions without any overt programming. MLAs are pressed into use in a wide range of apps such as email filtering and computer vision.



Understanding some of the main algorithms in this is quite useful. It would be more interesting if one knows what they are and where they fit. There are two ways to categorize the algorithms.

- The first being alignment of algorithms by virtue of their learning style.
- The second is alignment by virtue of their resemblance in their format or role (like aligning comparable animals together).

Though both ways of understanding are helpful, we concentrate on the second one to study variety of different algorithm types.

## 5 Proposed Study and Methodology

The India shape file is taken from the USGS, where the data or information is presented in both spatially and geographically. The geographic data is in the form of vector or raster data types. The data types have many formats like shape file, GeoJSON, GML, KML, GPX etc. Firstly, shape file is the general geospatial file type. All other commercial and open sources recognize shape file in GIS format. There are three required formats; SHP: is the feature geometry, SHX: is the shape index position, DBF: is the attribute data while PRJ, XML, SBN, and SBX are included optionally. Different MLAs are used for the classification of data. The following algorithms are used for identification of cities in India.

### 5.1 *K-NN Algorithm*

The algorithm names k-nearest neighbours (KNN) is a simple to use machine learning algorithm administered under supervision to solve categorization as well the regression problems. The KNN algorithm presumes that related things are present in shut proximity. As to say, related things are close to each other.

The very purpose of the k-Nearest Neighbours (KNN) algorithmic [16] rule is to use information during which the data points are differentiated into many different categories to envisage the categorization of a new sample point. This type is best understood through examples.

For illustration, we deem that each of the features in the training set as a dissimilar dimension in a certain space and take the value of the observation under consideration to be its coordinate in the aspect, so as to get a series of points in space. We then look at the similarity of any two points to be the gap amid them in that space below some suitable metric. During which the algorithmic rule decides which of the points from the training set are similar once selecting the category to foresee a new observation is chosen the k adjoining data points to the new observation, and to regard it as the most common class among them. Hence it is justified the name k Nearest Neighbours algorithmic rule. The algorithm (as described in [17, 18]) can be summarised as:



*Algorithm:*

- Step 1 In particular, a positive integer  $k$  is considered together with a new sample.
- Step 2  $k$  entries in our database that lay adjoining to the new sample are chosen.
- Step 3 The next step is to find the common categorization of the selected entries.
- Step 4 This is the categorization that is given to the new sample.

## 5.2 K-Means Clustering

Clustering is a expressive task that searches to identify consistent groups of objects depending on values of their attribute [19]. Mac Queen proposed K-Means algorithm in 1967 and is one of the frequently used clustering algorithms. This is deemed as one of the easiest unsupervised learning algorithms that classify the proposed data into  $k$  clusters so that intervals cluster add of squares is reduced.  $k$ -means clustering algorithm have a number of variants, which use an associate iterative scheme that functions well over a hard and fast range of clusters, using the following properties: Each class has a centre which acts as mean position of all the samples in that class. The algorithm can be summarised as

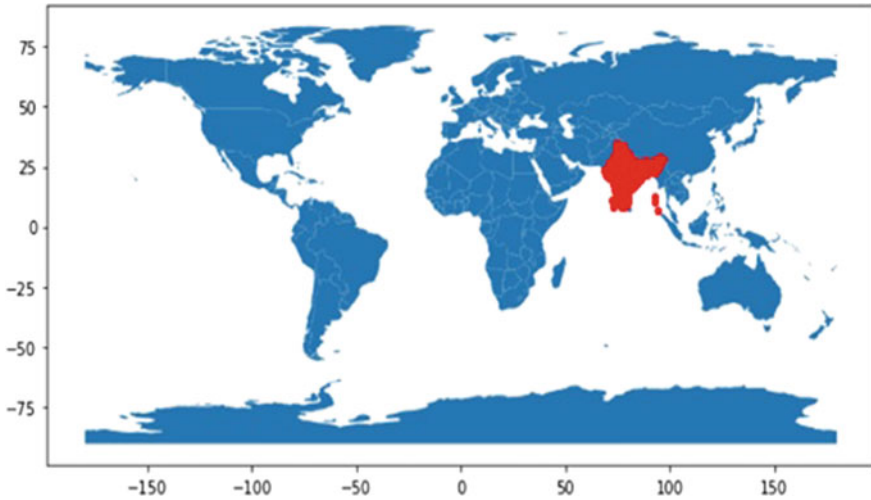
*Algorithm:*

- Step 1 Take the first three random group centroids into the 2D space.
- Step 2 Every object which has the closest centroids is assigned to the cluster.
- Step 3 The positions of the centroids is recalculated.
- Step 4 If the places of the centroids unaltered, proceed to the next step,  
Else go to Step 2.
- Step 5 End.

## 6 Experimental Analysis

Entire globe is taken to perform the work wherein India map is selected as shown in Fig. 5. India map shape file with every state which is shown in Fig. 6 is used. The shape file is converted into latitude and longitude information along with cities' data. This information is in the form of comma separated values (.CSV). The file contains the data of each city with their district, state and latitude and longitude data as shown in the Fig. 7. The data has 2340 and 18 fields are generated in the data file which can be used for the experimental analysis.

The file has the uncountable data which cannot read normally. So the information file is loaded in Mongoddb which is a Big data processing tool to store large data. The data stored in the Mongoddb is retrieved in python using mongo connector packages. In python, firstly we read the file from Mongoddb connector into python then applied the K-means clustering to plot the two-dimensional centroids by using the latitude



**Fig. 5** Visualization of India map in world map

and longitude coordinates as shown in Fig. 8. In the figure, only a few data can be marked to identify the places on a map to avoid the clumsiness of plot. On applying the k-nearest neighbour algorithm for the data file, we find the closest of every city in India and visualize the end points with bubbling every boundary coordinates as in Fig. 9. Adding every cluster and neighbour points to visualized map as shown in Fig. 10 from the entire India, Andhra Pradesh state is extracted through latitude and longitude of data as shown in the Fig. 11.

## 7 Conclusion

In this paper, globe is taken and selected India shape file is taken for identification of cities and the shape file is converted into GIS data file which is in the form of .CSV file format. The information contains the latitude, longitude, cities along with district and states of India country. The file is huge in size so firstly, uploaded in Big data tool called Mongoddb. Then Mongoddb is connected to python platform to retrieve the information. And then predicted India map in the world map, by k-means algorithm classified every city with two dimensional scatter plots with naming of every city. By k-NN algorithm, it is possible to predict every connected joint on the boundary in the form of visual image. Each clusters and neighbourhood data added to one another to form a visualized image of India. Also extracted is Andhra Pradesh state by pointing the locations of latitude and longitude from the GIS data file. Further we can classify each and every state to find cities and boundary and also find every cities land cover example: buildings, agriculture land, water floating, roads, railways etc., location of the particular land mark, particular area boundary and so on. Not only .CSV format

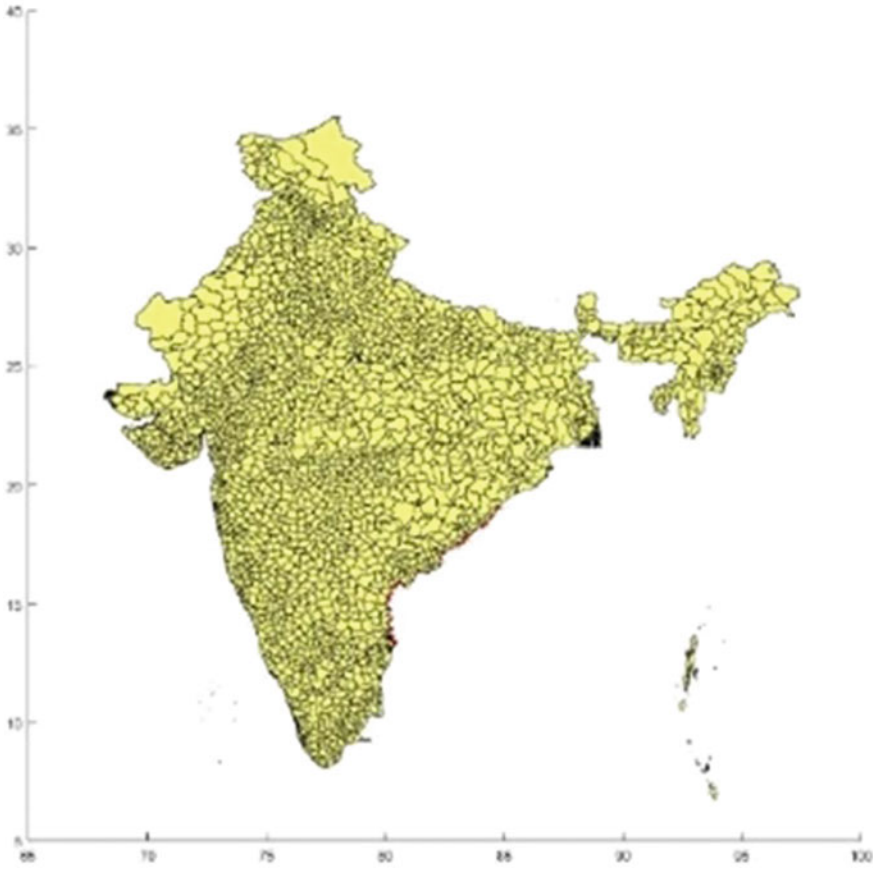


Fig. 6 India shape file

	X	Y	GID_0	NAME_0	GID_1	NAME_1	NL_NAME_1	GID_2	NAME_2	NL_NAME_2	GID_3	NAME_3	VARNAME_3	NL_NAM
0	93.809963	6.999320	IND	India	IND_1_1	Andaman and Nicobar	NaN	IND_1.1_1	Nicobar Islands	NaN	IND_1.1.1_1	n.a (2304)	NaN	
1	92.831696	12.611582	IND	India	IND_1_1	Andaman and Nicobar	NaN	IND_1.2_1	North and Middle Andaman	NaN	IND_1.2.1_1	n.a (2178)	NaN	
2	92.489004	10.704416	IND	India	IND_1_1	Andaman and Nicobar	NaN	IND_1.3_1	South Andaman	NaN	IND_1.3.1_1	n.a (2178)	NaN	
3	77.585558	14.693996	IND	India	IND_2_1	Andhra Pradesh	NaN	IND_2.1_1	Anantapur	NaN	IND_2.1.1_1	Anantapur	NaN	
4	77.601389	14.389240	IND	India	IND_2_1	Andhra Pradesh	NaN	IND_2.1_1	Anantapur	NaN	IND_2.1.2_1	Dharmavaram	NaN	

Fig. 7 File which has latitude and longitude data of each cities

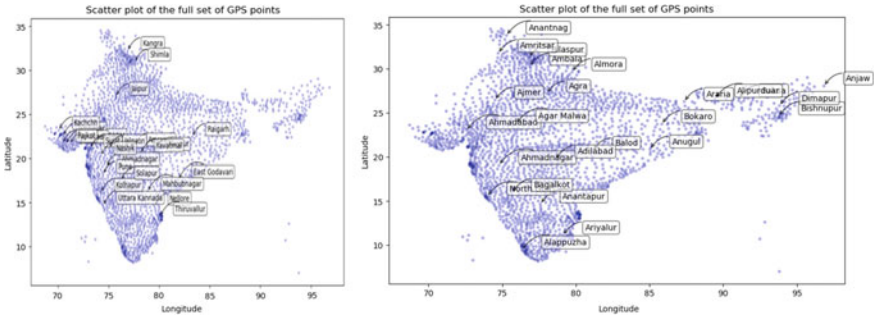


Fig. 8 K-means clustering for two dimensional scatter plot for India

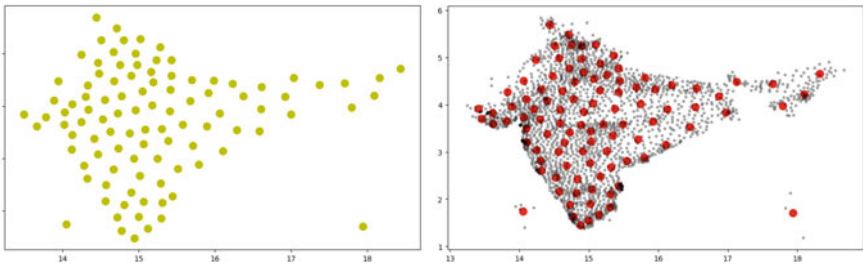


Fig. 9 K-NN algorithms for cities with boundary connected

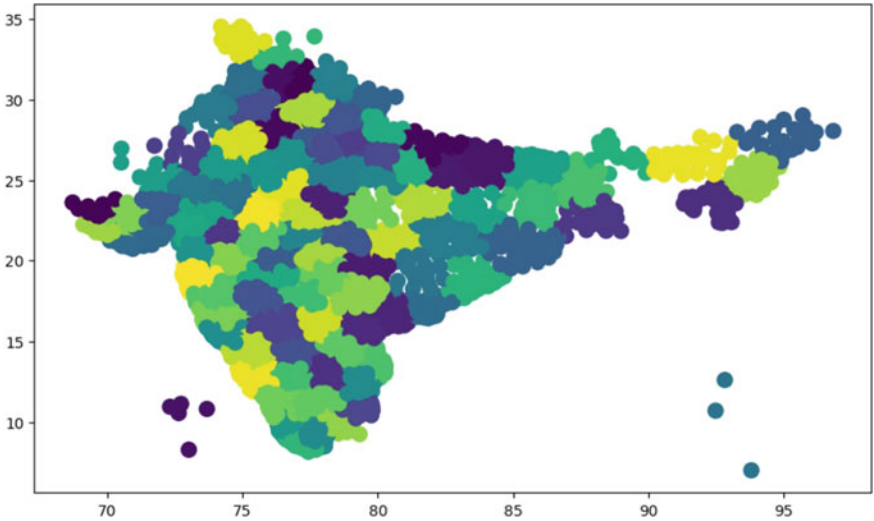
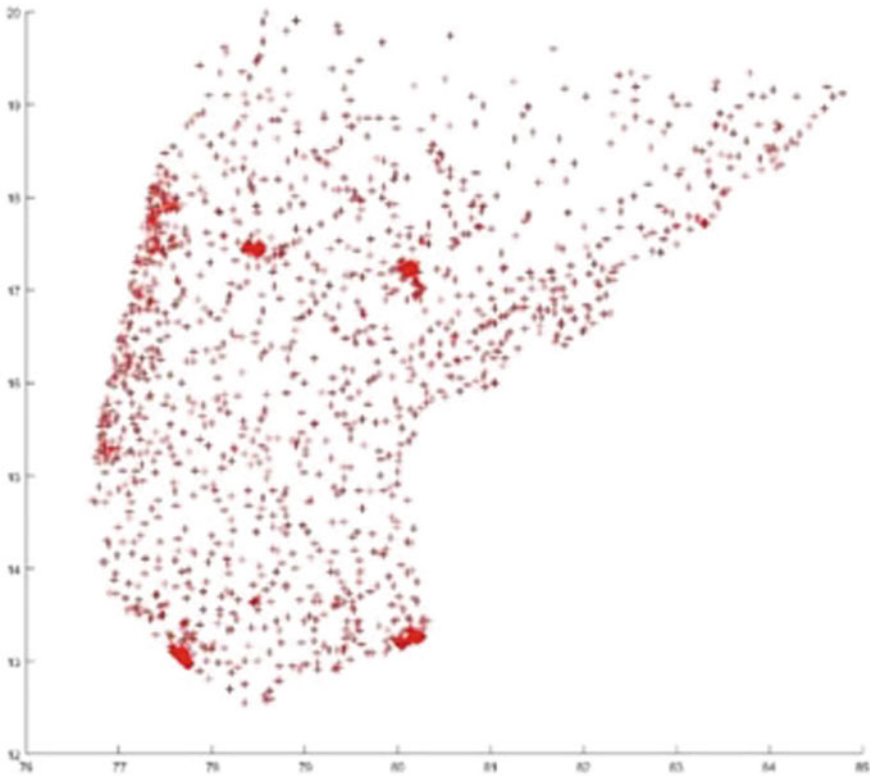


Fig. 10 Formed by adding clusters and neighbourhood



**Fig. 11** Extraction of Andhra Pradesh state from India by using latitude and longitude

file but also we can take JSON, GEOJSON, XML files and not only Mongoddb, FIG, HIVE tools of Big data is used for storing of Big data.

## References

1. T. Sarath, G. Nagalakshmi, An land cover fuzzy logic classification by maximum likelihood. *Int. J. Comput. Trends Technol. (IJCTT)* **13**(2) (2014) [arXiv:1407.4739](https://arxiv.org/abs/1407.4739)
2. G. Nagalakshmi, T. Sarath, An SVM fuzzy logic classification for land cover. *Int. J. Manag., Technol. Eng.* (2018). ISSN NO: 2249-7455
3. M. Sirish Kumar, B. Kavitha, S. Jyothi, G. Nagalakshmi, Land use/land cover of Tirupati area for agriculture land classification: a study. *Int. J. Eng. Res. Comput. Sci. Eng. (IJERCSE)* **5**(4) (2018)
4. T. Sarath, G. Nagalakshmi, S. Jyothi, A study on hyperspectral remote sensing classification. *Int. J. Comput. Appl.* (2015), in *International Conference on Information and Communication Technologies* (2014). doi: 0975-8887
5. K. SreeDivya, P. Bhargavi, S. Jyothi, Machine learning algorithms in big data analytics. *Int. J. Comput. Sci. Eng.* **6**(1), 63–70 (2018). E-ISSN. 2347-2693

6. J. Jeong, *Three Major Factors for a Successful Big Data Application* (National Information Society Agency, Seoul, 2012)
7. K. Song, *Understanding Society Through SOCIAL Metrics* (Daum Soft, Seoul, 2011)
8. B. Saha, D. Srivastava, Data quality: the other face of big data, in *Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE)*, pp. 1294–1297 (2014)
9. D. Zielstra, A. Zipf, A comparative study of proprietary geodata and volunteered geographic information for Germany, in *Proceedings of the 13th AGILE International Conference on Geographic Information Science* (2010)
10. L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, D. Keim, Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems, in *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 173–182 (2012). <https://doi.org/10.1109/vast.2012.6400554>
11. A. Tanuja, D. Swetha Ramana, Processing and analyzing big data using Hadoop. *Int. J. Comput. Sci. Eng.* **4.4**, 91–94 (2016)
12. X. Gao, J. Cai, Optimization analysis of urban function regional planning based on big data and GIS technology. *Techn. Bull.* **55**(11), 344–351 (2017)
13. B. Shneiderman, The big picture for big data: visualization. *Science* **343**(6172), 730 (2014). <https://doi.org/10.1126/science.343.6172.730-a>
14. The Flow of Human Migration [Internet] (2018). Available from: <http://www.public.tableau.com>. Accessed: 1 Sep 2018
15. D. Huffman, On the abuse of Chernoff faces. *Catastrophe* [Internet]. 2010. Available from <http://cartastrophe.wordpress.com/2010/06/16/on-theabuse-of-chernoff-faces>. Accessed: 2 Sep 2018
16. N.S. Altman, An introduction to kernel and Nearest-Neighbor nonparametric regression. *Am. Stat.* **46**(3), 175–185 (1992). <https://doi.org/10.1080/00031305.1992.10475879>
17. L. Kozma, *k Nearest Neighbours Algorithm* (Helsinki University of Technology, 2008). Available: <http://www.lkozma.net/knn2.pdf>
18. E. Mirkes, *KNN and Potential Energy (Applet)* (University of Leicester, 2011). Available: [www.math.le.ac.uk/people/ag153/homepage/KNN/KNN3.html](http://www.math.le.ac.uk/people/ag153/homepage/KNN/KNN3.html)
19. F. Maselli, G. Chirici, L. Bottai, P. Corona, M. Marchetti, Estimation of mediterranean forest attributes by the application of K-NN procedures to multi-temporal landsat ETM+Images. *Int. J. Rem. Sens.* **26**(17), 3781–3796 (2005). <https://doi.org/10.1080/01431160500166433>

# Secure Data Transfer Through Whirlpool—A Miyaguchi-Preneel Mode



Prasanna Mala Chelamkuri, E. G. Bhavya Reddy,  
and Annapurnaeswari Jonna

**Abstract** With increased usage of technology and internet in present scenario, there is huge need for data storage. According to internet world statistics the percentage of internet users are 100% of world's population. Where data sharing is the major operation performed in the internet with this vast usage, Security for the data became a major concern. And there is a huge need to protect the information we store on the internet. One of the approach is implementing hashing. To define hashing, it is the process of converting an input of any length into a fixed sized string of text using a mathematical function. One main advantage of using hashing is that they are unique. This paper presents one way collision resistant compressed 512 bit cryptographic hash function that works on a miyaguchi preneel mode.

**Keywords** Miyaguchi-preneel · Merkle-Damgård · Message digest · Compression function · Security · Data transfer · Hash function

## 1 Introduction

Now a days everything is getting digitalized, people are more often utilizing digitalized services. Since it was time saving and allow fast data transfer, although it is advantageous to use internet services for sharing information, there are some necessary issues with security [1]. Many security attacks have been developed for hacking data. So it is important to ensure security for the data before using internet. This paper

---

P. M. Chelamkuri (✉)

Department of Computer Science & Engineering, Jawaharlal Nehru Technological University,  
Kakinada, Andhra Pradesh, India  
e-mail: [prasannamala98@gmail.com](mailto:prasannamala98@gmail.com)

E. G. Bhavya Reddy · A. Jonna

Department of Computer Science & Engineering, School of Engineering and Technology,  
Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India  
e-mail: [bhavyareddy.eg@gmail.com](mailto:bhavyareddy.eg@gmail.com)

A. Jonna

e-mail: [jonnaannapurna@gmail.com](mailto:jonnaannapurna@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_11](https://doi.org/10.1007/978-3-030-46939-9_11)

presents an approach for data security, is through using a strong hashing technique i.e. Whirlpool, it is a block cipher based model that works on a Miyaguchi Preneel construction created by Vincent Rijmen (co-creator of Advanced Encryption Standard) and Paul S. L. M. Barreto. It was first written in the year 2000. This hashing formula is one of only two hash functions endorsed by NESSIE (National European Schemes for Signatures Integrity and Encryption) [2] also it was aided by ISO/IEC 10118-3 [3] is a specialized system for worldwide standardization of national bodies.

## 2 Related Work

Although there exists many hashing algorithms, out of all only few are providing better security to the data. Table 1 indicates the various hashes and their level of security.

**Table 1** Table indicating the different hashing algorithms

Hash	Number of bits	Passes broken		Author	Date launched
SHA-1	160	80	Yes	NSA	1995
SHA-2			None <sup>a</sup>	NSA	2000
SHA-256	256	64	None <sup>a</sup>	NSA	2000
SHA-384	384	80	None <sup>a</sup>	NSA	2000
SHA-512	512	80	None <sup>a</sup>	NSA	2000
MD2	128	1	Yes <sup>b</sup>	Ronald Rivest	1989
MD5	128	1	Yes	Ronald Rivest	1991
HAVAL	128		No	Yuliang Zheng, Josef Pieprzyk, Jennifer Seberry	1992
RIPEMD-320	320		No	Hans Dobbertin, Antoon Bosselaers, Bard Preneel	1996
GOST	64		No <sup>c</sup>	Soviet Union	1970s
Whirlpool	512		No <sup>d</sup>	Paulo Barreto, Vincent Rijmen	2001

<sup>a</sup>Although no attacks are reported people are doubtful about the security SHA-2 will provide. Because it is closely based up on the SHA-1 algorithm

<sup>b</sup>MD2 has been proved that it has some vulnerabilities and it is further concluded as not a secured cryptographic hash function

<sup>c</sup>GOST was developed and was used from 1970 by USSR, is not providing the security level required by ISO

<sup>d</sup>No attacks have been reported on earlier versions of whirlpool, the latest version of whirlpool is better in both software and hardware implementation



### 3 Implementation

#### 3.1 Design

This repeated hash structure was proposed by Merkle and Damgård, and it is used in almost all secure hash functions. This algorithm involves repeated use of round function which takes 2 inputs and produces 512 bit output. This cryptographic technique is a block cipher [4] based hash formula aimed to provide better protection to the data. Authors have declared that they will never take patent rights for this cryptographic technique and therefore it can be used for free of charge. There exists various versions of this cryptographic technique, due to some flaws identified in the earlier versions and making the significant changes to them, the latest version whirlpool is adopted. This W cipher is based on AES [5] and is designed to provide better implementation in both software and hardware that is both compact and exhibits better performance. In second version, a defect in the W cipher was found by Shirai and Shibutani that lowered the security of the algorithm, doing some significant changes solved this issue.

#### 3.2 Overview

Whirlpool [6] considers a message length of  $<2^{256}$  and produces a 512 bit message digest. The input is generated in 512 bit blocks. Figure 1 is the overall processing of

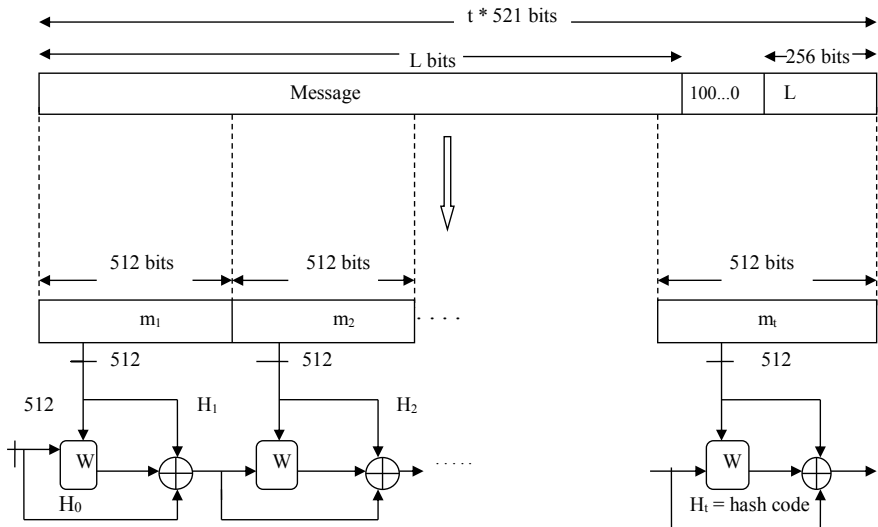


Fig. 1 A whirlpool structure overview

the message.

Here a message is needed to be padded, before being processed. The padding is the single 1 bit followed by the necessary numbers of 0 bits to make the length of the padding an odd multiple of 256 bits. After padding, a block of 256 bits is added to define the length of the original message. This block is treated as an unsigned number, after padding and adding the length field the augmented message size is an even multiple of 256 bits or a multiple of 512 bits. Whirlpool [6] creates a digest of 512 bits from a multiple 512 bit block message. The 512 bit digest  $H_0$  is initialized to all 0's. This value becomes the cipher key for encrypting the first block of the cipher text resulting from encrypting each block becomes the cipher key for the next block after being exclusive-ored with previous cipher key and the plain text block. The message digest is the final 512 bit cipher text produced after the last exclusive-or operation. Here the blocks of the cipher text are called the heart of the algorithm. Because the actual encryption starts here. Hence this block cipher  $W$  is the crucial part in encryption it undergoes some multiple round functions by performing continues encryption to the data and produces an 512 bit message digest. Here no matter how the message size is, it produces the same 512 bit hash value, hash values are generated for the no length string too. Whirlpool algorithm has an avalanche effect [7] that is even a small change in the input produces a significantly different output.

### 3.3 Whirlpool Block Cipher $W$

Whirlpool cipher is a non-feistel cipher like AES [5] that was mainly designed as a block cipher to be used in a hash algorithm [6]. The implementation of this hash function is very distinct than that of MD5 and SHA-1.

**Rounds:** whirlpool is a rounded cipher that uses 10 rounds. The block size and key size are 512 bits. The cipher uses eleven round keys each of 512 bits (Fig. 2).

$W$  is type of cipher that encrypts text by running blocks of a text through an algorithm by jumbling it. This is in contrast to a stream cipher that encrypts text to a one bit at a time, whirlpool uses this block cipher technique to encrypt the text by diving the unencrypted data in the form of blocks and executes each block. It go through some process called round function. This encryption process is crucial in this algorithm because the actual unencrypted data gets encrypted by undergoing round functions. The 4 round functions are

1. Substitution bytes
2. Shift column
3. Mix rows
4. Add round key.

The [6] Substitution byte function consists of a  $16 \times 16$  matrix called S-box that contains all possible values of 256 in 8-bit values. The leftmost 4 bits are considered as rows and the rightmost 4 bits are taken into consideration as columns that act as the index to the s-box to pick a different eight-bit value. For example 6th row

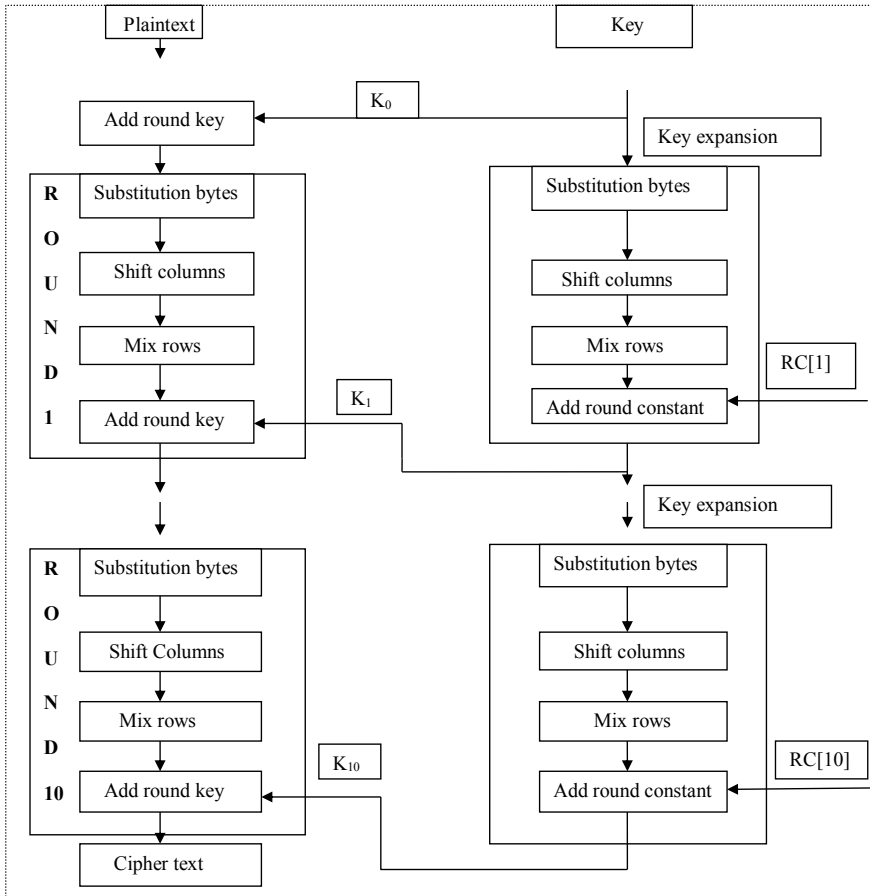


Fig. 2 Whirlpool block cipher W

4th column is BE and this BE can be written in the form of binary representation and leads to further implementation. The S-box can be generated using a recursive structure. Each containing two  $4 \times 4$  S-box separated by  $4 \times 4$  randomly generated box, each of the boxes map 4 bit input into a 4 bit output. The Shift column performs a circular downward shift of each column except the first column. For 2nd column performs 1 byte circular shift and for 3rd column performs 2 byte circular shift and so on. Here the shift column function serves as the permutation layer. The Mix rows functions performs diffusion layer it allows diffusion in each row individually, each byte of the row is mapped into a new value that is an  $8 \times 8$  bytes in that row this transformation can be defined as the matrix multiplication i.e.  $B = CA$  where A is the input matrix and B is the output matrix and C is the transformation matrix. C matrix is the predefined matrix. All the elements in the C are hexadecimal, here the addition and multiplication operations are performed in the  $GF(2^8)$  which is called Galois

field(named in the honor of Evariste Galois) [8] with the irreducible polynomial  $f(x) = x^8 + x^4 + x^3 + x^2 + 1$  i.e. 11D. In the Add round key layer the 512 bits of C state are bitwise XORed with the 512 bits of the round key. It is done byte by byte on the C state matrix.

### 3.4 Whirlpool Hashes

Whirlpool algorithm had go through two versions up to now from the year it started i.e. in 2000. People who would like to use whirlpool will be using the most recent version of whirlpool because it has better hardware and software implementation than earlier versions, and is also likely to be more secure. The 512 bit whirlpool hashes are commonly made as 128 digit sexadecimal numbers. Here the same ASCII byte input for the corresponding whirlpool hash versions produces different hash digest i.e. say “hey how do you do.” Produces a different hash value in Whirlpool 0, in Whirlpool T and in the Whirlpool.

Even a small change in the plaintext will result in distinct hash which will look completely different just like two unrelated random numbers [7]. Hash also exists for a zero length string.

### 3.5 AES Versus W

Though W is the modified version of AES there exists a huge difference between both algorithms (Table 2).

**Table 2** Table indicating the differences between the Whirlpool hash algorithm and AES (Advanced Encryption Standard)

	Whirlpool	AES
Block length	512 bits	128 bits
Key length	512 bits	128, 192 or 256 bits
Matrix input direction	Row	Column
Rounds	10	10,12 or 14
Key extension	Whirlpool round function	Special extension algorithm
$GF(2^8)$	$x^8 + x^4 + x^3 + x^2 + 1$	$x^8 + x^4 + x^3 + x + 1$
Round constant	From S-box	From $GF(2^8)$ 's $2^i$
Diffusion	Right multiply by a $8 \times 8$ matrix	Left multiply by a $4 \times 4$ matrix
Shift	Column	Row

## 4 Advantages

The proposed hash function mainly used for comparing two texts, checking data rectitude, detecting duplicates, and validation, pass code protection and file sharing. Here are some scenarios where proposed hash formula can be used

1. John wants to send a file to Tom and that file is highly confidential and John want to assure security while sending the file. One scenario is to handover the file directly to Tom then there is no point of digitalization. Another approach is to share via internet in a secured way then John uses a hash formula and generates a key for that file then John share the file and key to Tom. Now Tom receives both the file and the key. So now, Tom can compare both hashes. If they're the same, it means it's generated from the same file otherwise different. In this way Tom can verify if the file isn't in any way corrupted.
2. We can use this hash formula for storing pass codes, because hashes are one way functions i.e. they are not reversible once the pass code is hashed then that will be stored in the form of  $H(\text{"pass code"})$ . Whenever user login to any website, using some key exchange algorithm say Elgamal or Diffie-Helfman key exchange. User use that key to encrypt the pass code ( $x = e(\text{"pass code"})$ ) and send it to the server. The server, since it has the same key user used to encrypt the pass code, will perform some functions. Now the server performs the function  $H(\text{"pass code"})$  and compares it against the stored pass code hash it has associated with a username. If the hash generated from the value and user sent hash value matches with the one that server have, they allow access to the account, else no.
3. Using proposed hash technique without viewing the documents user can check the two documents for their equality. User can generate a hash value for both documents if the generated hash is same then the documents are same otherwise different.
4. This algorithm is used in checking an encrypted file last modified time. Because there are some encryption functions which never knew their last modified time like virtual device containers then that leads to uncertainty about the modifications done on the file. But using whirlpool due to avalanche effect [7] a small change in the file produces a different output.

## 5 Conclusions

Although it is not studied and tested much better, due to its robust scheme and compression function it is best in providing the security to the data. And it is based on AES, hence it is very resistant to attacks, because AES has proved to be resistant to attacks. Also reference implementations written in now are most widely used programming languages like C, Java that are available in the public domain for free of charge. There are some cryptographic programs that started using whirlpool in 2005 though now they were not in use but in 2005 they are most well known programs.

Whirlpool supports Recursive Hasher an open source command-line tool, which can calculate and verify whirlpool hash. Uses Ruby whirlpool library ironclad: a common list processing programming language which was a cryptography package containing a whirlpool implementation also ISO/IEC 10118-3 standard test vectors for the whirlpool hash from the NESSIE project manage C# implementation. Furthermore due its robust scheme (miyaguchi preneel) this algorithm can be implemented in various domains like banking, military, hospitals etc. also implementing this algorithm in intelligence agencies for sharing the confidential information provides security to the data, up to now there are no know attacks against whirlpool the latest version of whirlpool was design with faster execution and better characteristics, this can even be implemented in smart cards.

## References

1. [https://www.researchgate.net/publication/327799867\\_Introduction\\_to\\_Digitalization\\_Cases\\_How\\_Organizations\\_Rethink\\_Their\\_Business\\_for\\_the\\_Digital\\_Age](https://www.researchgate.net/publication/327799867_Introduction_to_Digitalization_Cases_How_Organizations_Rethink_Their_Business_for_the_Digital_Age)
2. [https://link.springer.com/content/pdf/10.1007%2F3-540-45664-3\\_21.pdf](https://link.springer.com/content/pdf/10.1007%2F3-540-45664-3_21.pdf)
3. [https://en.wikipedia.org/wiki/ISO/IEC\\_JTC\\_1](https://en.wikipedia.org/wiki/ISO/IEC_JTC_1)
4. [https://www.researchgate.net/publication/322294203\\_Comparative\\_Analysis\\_of\\_Block\\_Cipher\\_Modes\\_of\\_Operation](https://www.researchgate.net/publication/322294203_Comparative_Analysis_of_Block_Cipher_Modes_of_Operation)
5. <https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf>
6. <http://mercury.webster.edu/aleshunus/COSC%205130/N-Whirlpool.pdf>
7. [https://en.wikipedia.org/wiki/Avalanche\\_effect](https://en.wikipedia.org/wiki/Avalanche_effect)
8. <http://www.cs.utsa.edu/~wagner/laws/FFM.html>

# Frequent Item-Set Mining Using Lexicographical Sequential Tree Construction on Map Reduce Framework



P. Venkateswara Rao, D. Srinivasa Rao, and V. Sucharita

**Abstract** Frequent itemset mining is playing an important research role for many aspirants all over the world. In all the research aspects efficiency and scalability are main retrospects which are being improved in FIM which is intensive. In the current scenario, the implementation of a sequential growth algorithm on the big data map reduce framework the lexicographic sequential tree construction for the identification of the frequent itemsets using the lexicographical order over the databases of transaction without incorporating any extreme search methodologies. The result signifies a wide variety of large database executions to prove the execution of this methodology as an efficient and improved scalability of the methodology. Further the incorporation of this technique with other pattern mining is quite beneficial on big data.

**Keywords** Frequent itemset mining · Map reduce framework · Sequential growth · Lexicographical

## 1 Introduction

In the present days Information exploration has appeared as one of the major research domain in order to draw out implied and useful knowledge. This information is able to be understood by humans easily. Originally, this information removal was calculated and evaluated personally using mathematical methods. Consequently, there are so many popular technologies are developed in semi-automated data exploration.

---

P. Venkateswara Rao · V. Sucharita (✉)  
CSE Department, Narayana Engineering College, Gudur, India  
e-mail: [jesuchi78@gmail.com](mailto:jesuchi78@gmail.com)

P. Venkateswara Rao  
e-mail: [vrsairam23@gmail.com](mailto:vrsairam23@gmail.com)

D. Srinivasa Rao  
CSE Department, Lakireddy Bali Reddy College of Engineering, Mylavaram, India  
e-mail: [srinumtechse@gmail.com](mailto:srinumtechse@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_12](https://doi.org/10.1007/978-3-030-46939-9_12)

Therefore, automated data exploration methods were provided to synthesis knowledge effectively. In all the research aspects efficiency and scalability are main retrospects which are being improved in FIM which is intensive. In this paper a new MapReduce algorithm called sequence growth is implemented.

## 2 Related Work

R. Agarwal and R. Srikant has written a research paper on Frequent Item Set Mining (FIM) in the year 1995 [1]. It is a extension work of the research by the author 1994 [2]. In both the papers apriori algorithm were being introduced. From that time the FIM has grabbed attentions and it became important topic in research in data mining.

Eclat [3] uses another approach to mine frequent item sets called breadth first search approach. With the growth of huge data it is very difficult to execute the Association rule mining [4]. This problem was solved with big data. Introduction of MapReduce [5] has given solution to the distracted Association Rule Mining. The Map reduce framework can be implemented using many algorithms [6, 7]. There are several Apriori [8–10] for frequent Item set Mining. There is hybrid method of Apriori [11, 12] to adapt on the map reduce framework. From the distributed file system the Map Reduce programs will get input data [13, 14]. From Several studies it is clear that processing large datasets for apriori based algorithms is even more severe [15, 16].

## 3 Proposed Methodology

The proposed methodology consists of four phases incorporated as one for the implementation of the methodology. The first phase is where the lexicographic sequential tree construction for the identification of the frequent itemsets takes place. The next one is used for the pruning strategy which is called as lazy pruning is one of the ranked efficient pruning methodologies. The later phase is the implementation of the defined methodology on the mapreduce framework. Last phase gives adaptive processing in which we can incorporate even the other mining techniques like the pattern mining and all with respect to requirement. In the current scenario the implementation of a sequential growth algorithm on the big data map reduce framework is done. The lexicographic sequential tree construction is shown in Fig. 1 for the identification of the frequent itemsets using the lexicographical order over the databases of transaction without incorporating any extreme search methodologies. The processing of the algorithm in phase 1 is shown in Fig. 2 and phase 2 processing is shown in Fig. 3. The figures and algorithms have been taken from a scalable and effective frequent item set mining algorithm for big data based on Map Reduce framework [17].



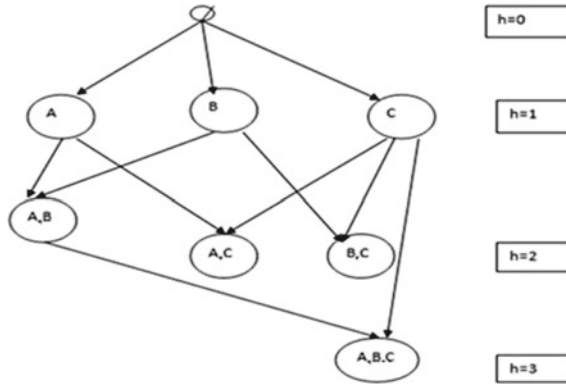


Fig. 1 Lexicographical tree structure example

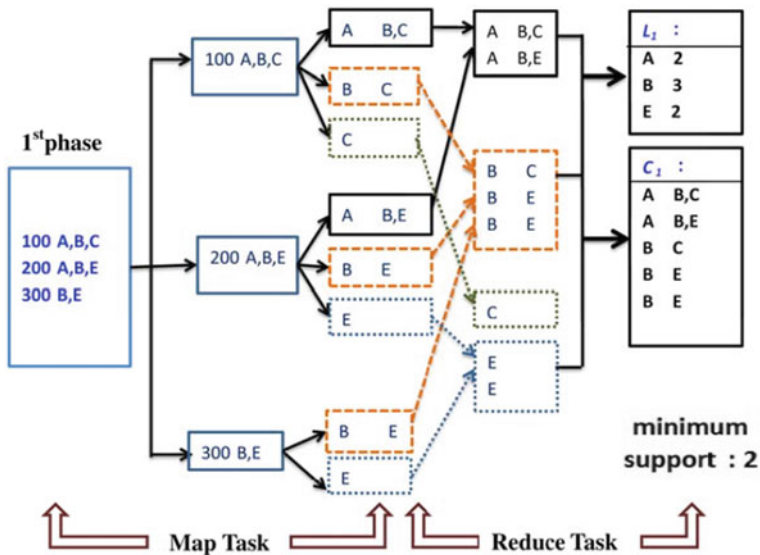


Fig. 2 Sequential growth Process example phase 1

### 4 Implementation

In the current scenario the implementation of a sequential growth algorithm [17] on the big data mapreduce framework is shown. The lexicographic sequential tree construction for the identification of the frequent itemsets using the lexicographical order over the databases of transaction without incorporating any extreme search methodologies is given. The pruning also plays an important role in making the methodology an efficient one. Each iteration has a map and a reduce task in sequence

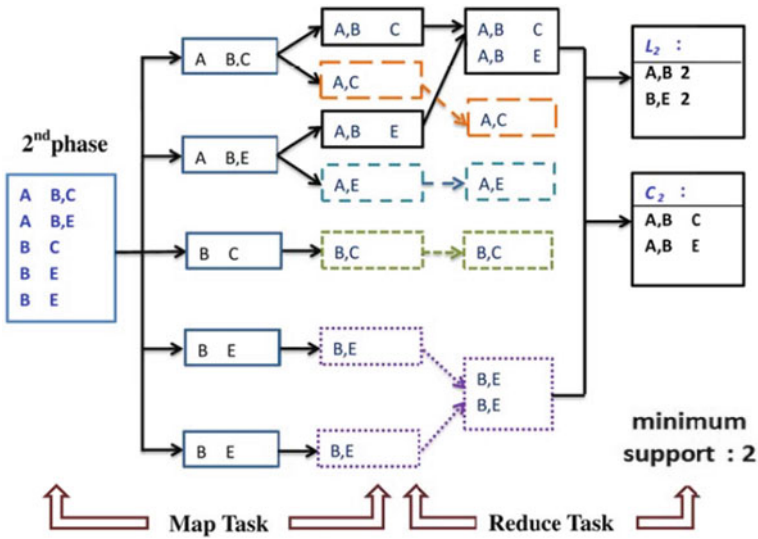


Fig. 3 Sequential growth Process example phase 2

growth algorithm. Sequence growth gives large – 1 item in the first step. Map reduce job is executed to produce length-k itemsets that occur frequently. Till the dataset of the output is empty the mapreduce iterations of the sequence growth continue. The steps of the sequence growth are shown in Algorithms 1–3.

**Sequential Growth—Algorithm 1:**

```

Input:  $S: \{ t \mid t \in S_i, t = \langle i_1, \dots, i_k \rangle \};$ 
        //A transaction database
         $\delta$ : integer; //minimum support threshold
Output:  $L: \{ p \mid p \in L, p = \langle key, value \rangle \};$ 
1: var[]  $L = \phi, C = \phi, T = \phi;$ 
2:  $S_i = \text{PartitionOf}(S);$  //  $S_i$  is a split of  $S$ 
3:  $(L, C) = \text{GenLarge1}(S_i, \delta);$ 
4: while  $(L \neq \phi)$  and  $(C \neq \phi)$  do
5:    $S_i = \text{PartitionOf}(C);$ 
6:    $C = \phi, T = \phi;$ 
7:    $(T, C) = \text{GenFrequentItemset}(S_i, \delta);$ 
8:    $L = \bigcup T;$ 
9: end while
    
```

**Algorithm 2: To generate 1 large items:**

**Input:**  $S_i: \{ t \mid t \in S_i, t = \langle i_1, \dots, i_k \rangle \};$   
 $\delta$ : integer; //minimum support threshold  
**Output:**  $L_1: \{ p \mid p \in L_1, p = \langle key, value \rangle \};$   
 $C_1: \{ c \mid c \in C_1, c = \langle key, value \rangle \};$

**Map Task** (*key*, *value*)

```

1: for each t in  $S_i$  do
2:   var[ ] itemlist;
3:   itemlist = t.split(",");
4:   for (k = 0; k < itemlist.length; k++) do
5:     key = itemlist[k];
6:     value = SuffixOf(itemlist[k]);
7:     Output(key, value);
8:   end for
9: end for

```

**Reduce Task** (*key*, *Value*[])

```

10: var sum = 0;
11: var[ ] subseq;
12: var[ ] str[2];
13: for each v in Value do
14:   sum++;
15:   subseq = ArraysCopyOf(subseq, sum);
16:   subseq[subseq.LastElement] = v;
17: end for
18: if (sum  $\geq$   $\delta$ ) then
19:   MultipleOutput(key, sum)  $\rightarrow$   $L_1$ ;
20:   for (k = 0; k < subseq.length; k++) do
21:     while (subseq[k].length  $\geq$  MaxMemory) do
22:       str = Split(subseq[k], MaxMemory);
23:       MultipleOutput(key, str[0])  $\rightarrow$   $C_1$ ;
24:       subseq[k] = str[1];
25:     end while
26:     MultipleOutput(key, subseq[k])  $\rightarrow$   $C_1$ ;
27:   end for
28: end if

```

**General FIM: Algorithm 3**

**Input:** :  $S_i: \{ t \mid t \in S_i, t = \langle \text{pattern}, \text{suffix} \rangle \};$

$\delta$ : integer; //minimum support threshold

**Output:** :  $L_k: \{ p \mid p \in L_k, p = \langle \text{key}, \text{value} \rangle \};$

$C_k: \{ c \mid c \in C_k, c = \langle \text{key}, \text{value} \rangle \};$

**Map Task** (*key* , *value*)

```

1: for each t in  $S_i$  do
2:   for (k = 0; k < itemlist.length; k++) do
3:     key = prefix.append(itemset[k]);
4:     value = SuffixOf(itemset[k]);
5:     Output(key, value);
6:   end for
7: end for

```

**Reduce Task** (*key* , *Value*[])

Same as the Algorithm 2

## 5 Evaluation and Results

The result signifies a wide variety of large database executions to prove the execution of this methodology as an efficient and improved scalability of the methodology by experimenting in Hadoop. Various transaction sizes are used for testing the scalability of the methodology. As the size of transaction increases there is a significant increase in the process time of the methodology. The capability of the methodology is useful even after the increase in the process time for the implementation on the large scale datasets. The implementation of this methodology on map reduction gives two fundamental tasks the map and reduce functionalities which follow the lexicographic sequential tree construction for the identification of the frequent itemsets. The pruning method incorporated in the proposed methodology greatly influences the efficiency of the methodology and decreases the intermediate data which further enhances the method efficiency. The experiments were also conducted to compare with other apriori based ones. The execution times were compared with different transaction lengths.

The iteration graphs depict that the proposed methodology is scalable and very efficient in the large scale implementation as shown in Fig. 4. Also the mapper receives an equally partitioned input which makes things even more simplified. The pruning makes the performance of the methodology increase a lot and is very important in the execution. The sequence growth with the datasets in millions is given in Fig. 5. The methodology supports the large scale transactions more efficiently and is very useful. It is also efficient in processing transactions lengths of high range as shown in Fig. 6. The results show that algorithm is very efficient in working with large data set when compared with Apriori and one phase.

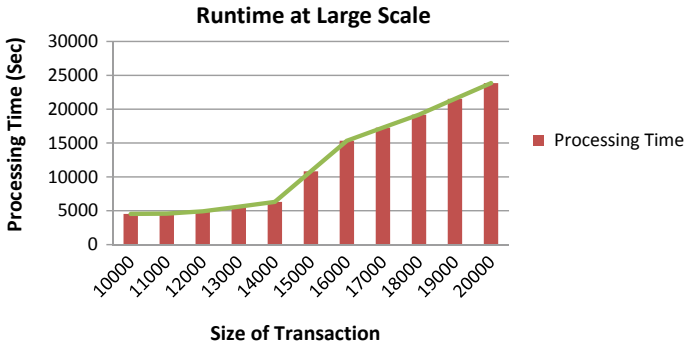


Fig. 4 Large scale transaction

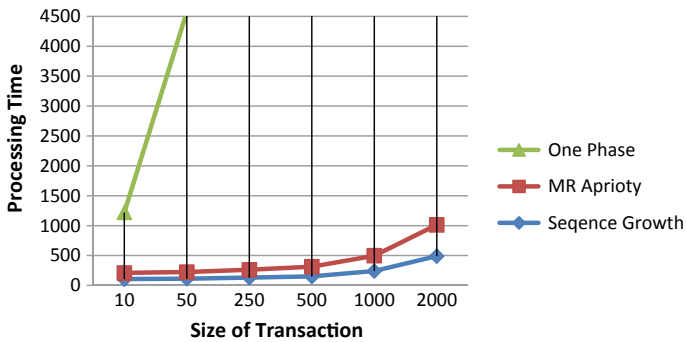


Fig. 5 Average transaction length

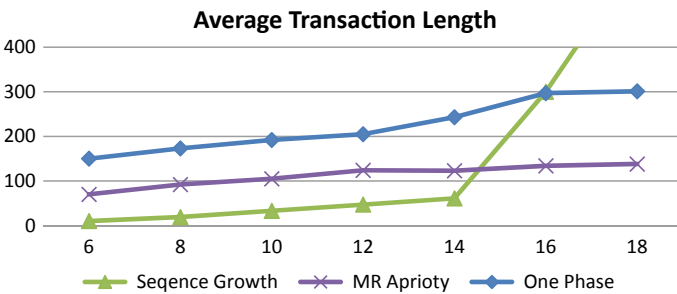


Fig. 6 Execution time comparison

## 6 Conclusion

The proposed methodology gives an effective procedure for the big data approach of extraction of frequent itemsets. As many distributed systems encountered the problem of intermediate data which is bypassed in the current methodology with the

help of the lexicographical sequential tree structure implementation. Also the pruning method incorporated in the proposed methodology greatly influences the efficiency of the methodology and decreases the intermediate data which further enhances the method efficiency. Without scanning the entire database repeatedly Sequence growth algorithm mines frequent itemsets.

## References

1. R. Agrawal, R. Srikant, Mining sequential patterns, in *Proceedings of the 11th International Conference on Data Engineering (ICDE95)* (1995)
2. R. Agrawal, S. Ramakrishnan, Fast algorithms for mining association rules, in *Proceedings of 20th International Conference on Very Large Data Bases, VLDB*, vol. 1215 (1994)
3. M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, New algorithms for fast discovery of association rules, in *Proceedings of the Third International Conference on Knowledge Discovery in Databases and Data Mining* (1997), pp. 283–286
4. D.C. Anastasiu, J. Iverson, S. Smith, G. Karypis, Big data frequent pattern mining, in *Frequent Pattern Mining* (Springer International Publishing, 2014), pp. 225–259
5. J. Dean, G. Sanjay, Map reduce: simplified data processing on large clusters. *Commun. ACM* **51**, 107–113 (2008)
6. C. Chen, C. Tseng, M. Chen, Highly scalable sequential pattern mining based on Map Reduce model on the cloud, in *2013 IEEE International Congress on Big Data (Big Data Congress)* (IEEE, 2013)
7. Z. Zhang, J. Genlin, T. Mengmeng, MREclat: an algorithm for parallel mining frequent item sets, in *2013 International Conference on Advanced Cloud and BigData (CBD)* (IEEE, 2013)
8. L. Li, M. Zhang, The strategy of mining association rule based on cloud computing, in *2011 International Conference on Business Computing and Global Informatization (BCGIN)* (IEEE, 2011), pp. 475–478
9. N. Li, L. Zeng, Q. He, Z. Shi, Parallel implementation of apriori algorithm based on Map Reduce, in *2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel and Distributed Computing (SNPD)* (IEEE, 2012)
10. O. Yahya, O. Hegazy, E. Ezat, An efficient implementation of Apriori algorithm based on Hadoop-Map Reduce model. *Int. J. Rev. Comput.* **12**, 59–67 (2012)
11. S. Hammoud, Map Reduce network enabled algorithms for classification based on association rules. Ph.D. thesis (2011)
12. S. Moens, E. Aksehirli, B. Goethals, Frequent item set mining for big data, in *2013 IEEE International Conference on Big Data* (IEEE, 2013), pp. 111–118
13. D. Huang, Y. Song, R. Routray, F. Qin, Smart cache: an optimized Map Reduce implementation of frequent item set mining. To appear in *IC2E* (2015)
14. A distributed Java-based file system for storing large volumes of data. <http://hortonworks.com/hadoop/hdfs/>
15. J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in *ACM SIGMOID Record*, vol. 29, No. 2 (ACM, 2000)
16. J. Pei et al., Mining sequential patterns by pattern growth: the fixspan approach. *IEEE Trans. Knowl. Data Eng.* **16**(11), 1424–1440 (2004)
17. Y. Liang, S. Wu, A scalable and effective frequent item set mining algorithm for big data based on Map Reduce framework, in *IEEE International Congress on Big Data* (2015)

# Deep Learning Based Recommender System Using Sentiment Analysis to Reform Indian Education



Jabeen Sultana, M. Usha Rani, and M. A. H. Farquad

**Abstract** Deep learning is a subset of machine learning, also known as hierarchical learning. It is based on artificial neural network with various stages of representative transforms. Deep neural networks have been applied in different applications like image processing, speech recognition; market-basket analysis and students' performance prediction to name a few. Now a day's education is not limited to only the classroom teaching but it goes beyond that like Online Education System, Web-based Education System, Seminars, Workshops, MOOC courses. It's a big challenge to extract sentiments from the huge data generated which is stored in the environments of Educational databases. Mining on educational databases can be done to extract the hidden sentiments of the students and their views about the education. Analyzing Students' sentiments and their learning behavior towards the course, difficulties faced, time spent for the course duration in learning the concepts and worries or fears of students like whether they may pass or fail the Final Exam is of prior importance these days in educational institutes. These factors play a dominant role in reforming education. Tweets are gathered from twitter database and found that the obtained are in unstructured form. Preprocessing methods were applied to clean the data set and later classified tweets based on sentiments into classes namely positive, negative and neutral. In this Paper, sentiments of students are analyzed which can be further considered while making reforms in education. In this paper Educational tweets are extracted from Twitter using twitter API and preprocessed. After Preprocessing, clean data is trained and a Model is attained, on this test data is applied. Results are evaluated on few parameters like Balanced accuracy, Sensitivity and Specificity; Prevalence and Detection rate and found that deep learning technique achieves high performance.

---

J. Sultana (✉) · M. Usha Rani  
Department of Computer Science, Sri Padmavathi Maha Vishwavidyalayam, Tirupati, India  
e-mail: [jabeens02@gmail.com](mailto:jabeens02@gmail.com)

M. Usha Rani  
e-mail: [musha\\_rohan@yahoo.com](mailto:musha_rohan@yahoo.com)

M. A. H. Farquad  
INI Labs, Waterloo, ON, Canada  
e-mail: [farquadonline@gmail.com](mailto:farquadonline@gmail.com)

**Keywords** Deep learning · Classification · MLP · Decision Trees (DTREE) · Naïve Bayes Tree (NBTREE) · SVM and Twitter

## 1 Introduction

In the modern times, education has shifted towards online to meet the needs of various categories of students as education system is one of the important parts for the development of any country. At the present time, heaps of data are composed in educational databases, but it remains unused. Educational data mining alarms through emerging ways aimed at determining knowledge from huge generated data that comes from educational domain. To obtain the aids from complex data, leading tools and technologies are essential. Data mining stands as evolving prevailing tool for analysis and prediction. Classification techniques have been successfully used in health domain, real-estate assessment, and intrusion detection and educational sectors. It is very useful in mining and analyzing educational sentiments to enhance student's performance and to make effective reforms in education.

Education system in any part of the world, basically considers norms for eligibility to get enrolled in an institute for a particular program with time duration as important factors in making educational reforms. Heaps of data is generated from educational institutes since everything is going online. Any institution which makes effective use of mining the educational data finds unique mode of enlightening student's performance, achievement level and attractiveness towards the particular program or course. This may definitely help in improving reforms like the excellence in offering education, more student intake, advising. Significant techniques in mining educational data are deep neural nets, classification, association rule mining and clustering. Supervised learning takes place in classification approach in which students are grouped into identified classes [1]. Rules for classification can be distinguished from data known as training data and further tested for the remaining data [2]. The classification system is evaluated for the effect of the rule's reliability on the test data set on few parameters.

Here, we suggest a classification model for student's sentiments prediction model using and deep learning approach on Indian educational tweets and compared with Classification approach. Deep Learning approach is used and compared with some techniques of data mining to predict student's sentiments in participating or interested to suggest while making educational reforms. MLP [3] which is a Deep Learning method compared with other set of classifiers like, Decision Tree [4], and Naïve Bayes Tree [5]. In addition to this, we evaluated these models by comparing the performance on various parameters like Balanced Accuracy, Specificity and Sensitivity in order to optimize and select the best Model. The obtained results reveal that MLP achieves best accuracy comparing with the other classifiers results.



## 2 Literature Survey

As most of us are using digital devices by affording data connections, lots and lots of data is generated online from different sites of social media like Instagram, Facebook, Twitter and LinkedIn., let us consider one leading application of social media, Twitter. It is used by different age groups of people all over the world to express their sentiments on diverse range of subjects, subtopics. Users may share their sentiments towards Politics, industries, educational institutions, marketing, security and awareness and much more. Lots of data in terms of petabytes is produced every day from twitter alone. So the need of preprocessing the data in order to make some meaningful sense arises in order to make better decisions. Data processing, also termed as text mining is widely used to extract meaningful information from lots of data, which is unstructured by nature. The extracted meaningful information or a proper data form will help to analyze the data in an effective way and make good decisions [6].

Sentiment analysis has gained huge attention and popularity, promising area for research. This has been gained attention since various social networks generated huge amount of big textual data from these networks and other information centric applications [7]. Sentiment analysis on customers' feedbacks for the online products wholesaled was performed using attributes based feedback summarization system. Natural Language Processing methods were used as they possess the ability to understand human language and their sentiments. Product features were mined and found the opinions to be positive or negative and then decision was made regarding the sentiments [8].

Furthermore, a model was proposed based on student feedbacks regarding their performance of teachers in Spanish to analyze the sentiments of students [9]. Innovative technologies for education are emerging and online education is becoming common and creates interest among the students. Enhances learning behavior among the students by offering various courses and scheduling the timings for the course based on student's flexibility. The big data collected from online databases like MOOCs desires to be huge, not in its instances but with more attributes and information on learners' cognitive and desires to be composed of automatic circumstances accumulating rightness and completion rates. This more detailed enunciation supports students learning process in analyzing the data from black box approach. Fine grain data is used by data-driven learner model approach that is considered and improved by using principles from cognition [10].

Latest novelties in e-learning lead to start courses online at numerous levels, over turned classroom teaching, business training at corporate sectors resulted in thought-provoking problems about SLN [11, 12]. Information exchange among individuals is broadly highlighted and intensely affected by sentiments of the utterers. individuals change their replies grounded on the activities of their partner discussions in a definite sensitive manner. Their sentiments are positive if they are happy, sad or upset if they are not happy and rude when they are in angry mood [13]. Traditional Deep

Learning approach was used to predict student's sentiments on educational data and was compared with machine learning approaches. It was analyzed that MLP obtained the optimal results [14].

Comparative analysis of student's performance on educational data was carried out and observed that SVM and MLP gave optimal results compared to decision tree and Naïve Bayes [15]. Sultana et al. [16] suggested a performance prediction model for student's using deep learning and data mining methods students' performance based on student's learning behavior The model was evaluated in different classifiers; Naïve Bayesian, Deep Neural Network and Random Forest. Deep Neural Network outstands with the rest of others in performance. Also, a survey on some deep learning applications was conducted signifying different uses of Natural Language processing, text mining, automatic navigation systems, speech recognition [17].

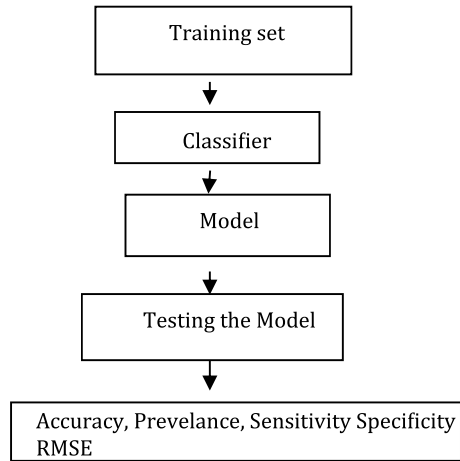
Basing the above literature reviews, we suggest to analyze the sentiments from Indian education tweets and classify the tweets based on polarity scores. Best techniques to classify the tweets with balanced accuracy is considered among all the techniques based on few parameters.

### 3 Data Collection and Preprocessing

In this paper, the educational tweets are collected from Twitter using twitter API developers account [18]. Twitter offers users to showcase their sentiments online from any device with Internet connection. The dataset consists of 350 instances with six features comprising educational tweets of India. The tweets are classified into three classes based on their polarity score of sentiments using R Package in windows systems.

### 4 Suggested Methodology

In this research, we have pre-processed data using preprocessing techniques like removal of hash tags, punctuations, quoted text, URLS, stop words etc. The Experiments are carried out on the preprocessed educational tweets collected from the twitter repository. Later, the obtained results were checked using confusion matrix. Balanced accuracy, Sensitivity, specificity is obtained using this. Sensitivity means the proposition of properly identified positive cases, specificity is calculated from here.

**Fig. 1** Methodology

### 4.1 Frame Work Demonstrating Suggested Methodology

The Suggested Methodology follows three stages

1. Data set is spitted into training and testing data. Classifiers like DTREE, NBTREE, MLP and SVM are selected and trained leading to a Model.
2. Test data is imparted on the Model and results are analyzed.
3. The obtained results at this stage are checked on parameters like Balanced Accuracy, Prevalance, detection rate, detection prevalence, sensitivity and specificity.

The framework illustrated here précises the suggested framework for analyzing educational sentiments of people in India in reforming education.

The Fig. 1 is showing the suggested Framework.

### 4.2 Methods Used

1. MLP: The principle method for training multilayer perceptrons is the back propagation algorithm (including its variants). In the MLP neural network, each node handles the amount of biased inputs and goes through this activation level to create a transfer function. The most common activation functions in MLP are logistic and hyperbolic tangent sigmoid functions [19].
2. Decision Trees: They are an integral part of [20], an integral part of ‘Machine Learning. C4.5 uses the split-and-conquer method to grow decision trees. These trees start at the root of the tree and go to its leaf nodes. The J48 algorithm that uses Decision Tree implementation is used in the experiments reported here. DT is widely accepted in decision-making systems and is used because of its human

understandable structure. A test item for the class label starts from the root of the tree and moves through it to the leaf node, which provides the classification of the instance.

3. Naïve Bayes Tree: Naive Bayes Decision Tree (NBTREE) is similar to DT except at the leaves. NBTREE is a hybrid of decision tree classification and naive-Bayes classification. The Bayes rule is used to calculate the probabilities of each class using the given examples. Every attribute value at a given label requires an estimate of the conditional probability. Classification at leaf nodes is done by NB classifiers. Compared to DT, the NBTREE potential is present at each node and the total probability is not greater than unity [21].
4. SVM: Support Vector Machine (SVM) is a universal structured learning method based on statistical learning theory [22]. SVMs are an inductive machine learning technique based on structured risk minimization, which classifies performances by minimizing real error and building an N-dimensional hyperplane that divides the data into two categories. The main goal of SVM is to find the right training hyperplane that accurately classifies data points and separates the two class points as much as possible, minimizing the risk of misclassifying training models and missing test models. SVM models are closely associated with neural networks.

Confusion Matrix: This is based on class labels, where the comparison is made between the actual class labels and the class labels classified by the classifiers. Word2Vec vectors as inputs to convolutional neural networks and has increased the accuracy of sentiment classification.

## 5 Result Considerations

In this section, we analyze the results.

The results are evaluated in terms of Balanced Accuracy, Specificity, sensitivity, Prevalance, Detection rate, and Detection Prevalance. A comparison was drawn among data mining techniques like Support vector machine, Decision tree and Naïve

**Table 1** The below table describes the results of different classifiers used here and they are Naive Bayes, DTREE, MLP and SVM

Performance measures	Naïve Bayes	Decision tree	SVM	MLP
Balanced accuracy	0.67	0.57	0.70	0.90
Specificity	0.66	0.70	0.75	0.82
Sensitivity	0.68	0.95	0.77	0.90
Prevalance	0.1	0.12	0.08	0.18
Detection rate	0.01	0.08	0.06	0.12
Detection prevalence	0.10	0.86	0.24	0.41

Bayes with respect to deep learning technique i.e., MLP. MLP outperformed in classifying educational tweets with high balanced accuracy of 90% followed by SVM., Naïve Bayes and decision tree. Other parameters like Sensitivity, Specificity, Prevalance, Detection rate and detection prevalence were considered for analyzing the best performance and it was found that MLP gave optimal results followed by the rest of the classifiers.

## 6 Conclusion and Future Directions

Mining techniques helps to improve and efficiently analyze sentiment analysis of twitter data in the field of education system as they generate bulk quantities of teaching and learning documents about in particular but not limited too. Twitter data we used here contains hidden sentiments of users towards education in India. This extracted sentiments information helps to improve educational norms and regulations in educational institutes. We suggested good performance yielding deep learning and data mining techniques. The sentiments of student's and people towards education system in India is predicted by training the classifiers and the obtained model is evaluated by set of classifiers, namely; MLP, DT, Naïve Bayes and SVM. Training and testing is accomplished.

The results specify that the deep learning method gives overall good performance in terms of classification accuracy, sensitivity and specificity and predictive rate. However, DT gives consistently fewer rules and therefore the perceptual rules are superior. From the results, the accuracy, sensitivity and specificity are lost in the process of retrieving the knowledge gained by DT, NBTree. Although DT is generally a superior classification compared to others, the knowledge that DT acquires is transparent, lacking accuracy. Therefore, it is advisable to use the MLP-effective learning method directly to make educational changes in the country. The scope of future work for this paper is to work on the rule extraction part and accuracy in DT, Naïve Bayes.

## References

1. A. Merceron, K. Yacef, Educational data mining: a case study, in *International Conference on Artificial Intelligence in Education AIED, Amsterdam* (IOS Press, 2005), pp. 467–474
2. I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan-Kaufmann Series of Data Management Systems (Elsevier, San Francisco, 2005)
3. S.K. Pal, S.K. Mitra, Multi-layer perceptron, fuzzy sets and classification. *IEEE Trans. Neural Netw.* **3**(5) (1992)
4. M.M. Quadri, N.V. Kalyankar, Drop out feature of student data for academic performance using decision tree techniques. *Glob. J. Comput. Sci. Technol.* **10**(2) (2010)
5. N.T.N. Hien, P. Haddawy, A decision support system for evaluating international student applications, in *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, FIE'07. 37th Annual* (IEEE, 2007), pp. F2A-1

6. N. Zhong, L. Yuefen, W. Sheng-Tang, Effective pattern discovery for text mining. *IEEE Trans. Knowl. Data Eng.* **24**(1) (2012)
7. S. Suprajha, C. Yogitha, J. Archita, H.S. Guru Prasad, A study on sentiment analysis using tweeter data. *Int. J. Innov. Res. Sci. Technol.* **1**(9) (2015)
8. M. Hu, B. Liu, Mining and summarizing customer reviews, in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA* (2004), pp. 168–177
9. T. Patel, J. Undaiva, A. Patel, Sentiment analysis of parents feedback for educational institutes. *Int. J. Innov. Emerg. Res. Eng.* **2**(3), 75–78 (2015)
10. G.G. Esparza, A.P. Diaz, J.C. Recih, C.A.D. Luna, J. Ponce, Proposal of a sentiment analysis model in tweets for improvement of the teaching-learning process in the classroom using a corpus of subjectivity. *Int. J. Comb. Optim. Probl. Inform.* **7**(2), 22–34 (2016)
11. M. Opuszko, J. Ruhland, Classification analysis in complex online social networks using semantic web technologies, in *IEEE Computer Society, International Conference on Advances in Social Networks Analysis and Mining* (2012), pp. 1032–1039
12. E. Maleki, A. Rezaei, M.B. Behrouz, Comparison of classification methods based on the type of attributes and sample size. *J. Convergence Inf. Technol.* **4**(3), 94–102 (2009)
13. M. Hasan, E.A., Rundensteiner, E. Agu, *EMOTEX: Detecting Emotions in Twitter Messages* (Academy of Science and Engineering, USA, ASE, 2014)
14. J. Sultana, N. Sultana, K. Yadav, F. Alfayez, Prediction of sentiment analysis on educational data based on deep learning approach, in *Proceedings of 21st Saudi Computer Society National Computer Conference (NCC)* (2018), p. 1
15. J. Sultana, M. Usha, M.A.H. Farquad, An efficient deep learning method to predict students performance, in *Higher Education Quality Assurance and Enhancement* (Rishi Educational Society Book Series, 2018). ISBN 978-81-936838-0-4
16. J. Sultana, M. Usha, M.A.H. Farquad, Student's performance prediction using deep learning and data mining methods. *Int. J. Recent Technol. Eng. (IJRTE)* (1S4), 1018–1021 (2019). ISSN: 2277-3878 (Blue Eyes Intelligence Engineering & Sciences Publication)
17. J. Sultana, M. Usha, M.A.H. Farquad, An extensive survey on some deep learning applications, in *Emerging Research in Data Engineering Systems and Computer Communications*. Advances in Intelligent Systems and Computing Series, vol. 1054. *Proceedings of CCODE* (2019), 978-981-15-0134-0
18. <https://dev.twitter.com/streaming/overview>
19. W.H. Delashmit, T. Michael, Recent developments in multilayer perceptron neural networks, in *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC*, vol. 699
20. J.R. Quinlan, *C4.5 Programs for Machine Learning* (Morgan Kaufmann, 1993)
21. R. Kohavi, Scaling up the accuracy of Naïve-Bayes classifiers: a decision tree hybrid, in *Proceedings of KDD-96, Portland, USA* (1996), pp. 202–207
22. V.N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn. (Springer, New York, 1998)

# An Analysis of In Vitro Antioxidant and Anti-inflammatory Activities of *Mucuna pruriens* (Leaves) and *Allium sativum* (Bulbs)



Bysani Jagannatha Divya, Bukke Suman, Mallepogu Venkataswamy, Kalla Chandra Mouli, and Kedam Thyaga Raju

**Abstract** The anti-inflammatory steroidal and non-steroidal drugs are commonly used to treat inflammation. Be that as it may, with a number of side effects, these are disabled. We have therefore selected two plants, i.e., *Mucuna pruriens* and *Allium sativum*, which have various therapeutic agents used in traditional medicines and Ayurveda. The objective of this investigation was to test in vitro anti-inflammatory activities of *Mucuna pruriens* leaves and *Allium sativum* bulb extracts of both either alone or in combination by evaluating inhibition of cyclooxygenase and with determination of protein denaturation. The successive solvent extraction of shade dried leaves and peeled dry bulbs was performed for both plants with solvents of increased polarity. Then the two plants' in vitro antioxidant assay was evaluated by estimating DPPH, H<sub>2</sub>O<sub>2</sub>, NO, ABTS and ascorbic acid as standard at various concentrations. The methanol and combined methanol extracts from the two plants showed significant antioxidant activity on DPPH, H<sub>2</sub>O<sub>2</sub>, NO and ABTS among the four extracts. In addition, the cyclooxygenase inhibition (76.23%) and protein denaturation (65%) were also tested for anti-inflammatory activity, which revealed that these two plants had strong antioxidant and anti-inflammatory principles.

**Keywords** Non steroidal · Anti-inflammatory · Denaturation · Cyclooxygenase · Anti oxidant

---

B. J. Divya (✉) · B. Suman · M. Venkataswamy · K. T. Raju  
Department of Biochemistry, Sri Venkateswara University, Tirupati,  
Andhra Pradesh 517502, India  
e-mail: [bysanidivya@gmail.com](mailto:bysanidivya@gmail.com)

K. T. Raju  
e-mail: [thyagarajuk\\_1999@yahoo.com](mailto:thyagarajuk_1999@yahoo.com)

K. C. Mouli  
Department of Botany, Sri Venkateswara University, Tirupati,  
Andhra Pradesh 517502, India

## 1 Introduction

Since times prehistoric, medicinal plants have been found and utilized in practically all cultures as a source of potential new medications. The boundless utilization of herbal remedies and herbal preparations, as those depicted in antiquated books, for example, Bible and the Vedas, and got from ordinarily utilized traditional herbs and medicinal plants, has been suggested the presence of herbal products with therapeutic properties [1]. Plants are the primary sources of therapeutic agents for treatment of different sicknesses and therapeutics against diseases since ancient time. The natural drugs have contributed in the advancement of modern medicine and their dynamic therapeutic principles [2]. The medicinal plants serve as a model to develop more effective and less harmful prescriptions utilizing their secondary metabolites, for example, alkaloids, flavonoids, phenols, saponins, sterols and so on to create pharmacologically active principles that may demonstrate independently, additively or in combination to improve health [3, 4].

Within traditional medicine, the use of these compounds is common and are the significant source of natural antioxidants that could lead to the development of novel drugs [5], ethno-medicines with strong antioxidant properties [6], and therapeutic potential for free radical related disorders [7].

Inflammation is a complex process that is often associated with pain, swelling, and includes events such as vascular permeability, protein denaturation, and membrane alteration. Denaturation of proteins is a process where proteins lose their secondary and tertiary structures by adding external pressure or compounds like strong acid or base, concentrated inorganic salt, organic solvent or water. Once denatured, the majority of biological proteins lose their function. Protein denaturation is a well known cause of inflammation.

The ability of plant extract to prevent protein denaturation was studied as part of the investigation into the mechanism of its anti-inflammatory activity [8, 9].

Inflammation processes require a cascade of events in which arachidonic acid metabolism plays a significant role. One of the mechanisms is that prostaglandins (PG) and thromboxane A<sub>2</sub>, which are essential biologically active mediators in a number of inflammatory events, can be metabolized by Cyclooxygenase (COX). Arachidonic acid is cleaved from membrane phospholipids when properly activated by neutrophils and can be converted to prostaglandins through the COX pathway. COX inhibition leads to a decrease in PG production, such a drug would have the potential to provide anti-inflammatory and analgesic effects with a decrease in the gastrointestinal side effects.

The primary sources of naturally occurring antioxidants are whole grains, fruits and vegetables [10]. Most secondary metabolites containing antioxidants are phenolic acids, polyphenols and flavonoids. In the above context, two plants were chosen for our study, *Mucuna pruriens* and *Allium sativum*.

*Mucuna pruriens* is part of the *Fabaceae* family, commonly referred to as a cowage plant. It is a common medicinal plant in India that has been used since ancient times. It is typically found in tropical regions and is used for various purposes in traditional



medicine in many countries. In the traditional medicine system in India and West Africa, all parts of *Mucuna pruriens* have valuable medicinal properties [11].

*Allium sativum*, commonly known as garlic, which belongs to the *Amaryllidaceae* family, is another medicinal plant selected for study. It is one of the earliest known plants of traditional medicine [12]. Garlic and its preparations have been commonly used for health benefits as a result of various research reports over the past decade [13].

Therefore considering the importance of plant products, these two plants, *Mucuna pruriens* and *Allium sativum* have been selected to study in vitro antioxidant and anti-inflammatory activities and to examine the secondary metabolites in the regulation. The combination of these two species in therapeutics is not being discussed. Common uses show that both plants are rich in anti-inflammatory source of antioxidant activity. The following procedures for analyzing the products of these crops have been introduced.

## 2 Materials and Methods

### 2.1 Collection of Plant Material and Preparation of Extracts

The fresh leaves of *M. pruriens* were collected from Seshachalam hills of Eastern Ghats (Tirumala Hills) in Andhra Pradesh, India. The fresh bulb cloves were bought from the local Tirupati market, Chittoor district of Andhra Pradesh. The Plant Taxonomist of Department of Botany, S.V. University, Tirupati has described and authenticated both the *Mucuna pruriens* (L) leaves and *Allium sativum* bulb cloves with voucher specimen numbers SVUBH/592 and SVUBH/1123 respectively. The fresh leaves of the *Mucuna pruriens* and peeled *Allium sativum* bulb cloves were dried in shade and milled with a mechanical grinder to fine powder. The powdered material was macerated separately by hexane, aqueous (water), methanol and ethyl acetate. The extract was then filtered with filter paper (Whatmann paper) and then using rota evaporator at 40 °C, the filtrate was reduced under pressure. The resulting concentrate was a dark molten mass layered on aluminium foil and freeze dried for further use.

### 2.2 In Vitro Antioxidant Activity

#### 2,2-diphenyl-1-picryl hydrazyl (DPPH) free radical scavenging assay

Free radical scavenging activity by DPPH was assayed according to the method reported by Burits and Bucar [14] with minor modifications.

### **Hydrogen peroxide scavenging assay**

The hydrogen peroxide scavenging assay was done according to the method reported by Vijayabhaskaran with some modifications [15].

### **Nitric oxide radical scavenging assay**

Nitric oxide (NO) radical scavenging activity was assayed according to the method reported by Garrat with slight modifications [16].

### **2, 2'-azino-bis (3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) radical scavenging Assay**

ABTS radical cation decolorization assay was used to determine free radical scavenging activity of plant samples according to the method reported by Pellegrini with slight modifications [17].

## **2.3 *In Vitro* Anti-inflammatory Activity**

### **Inhibition of protein denaturation**

Inhibition of protein denaturation was assayed according to the method reported by Sakat with slight modifications [18].

### **Separation of White Blood Cells**

White blood cells (WBCs) were isolated from blood according to the method reported by Arne [19].

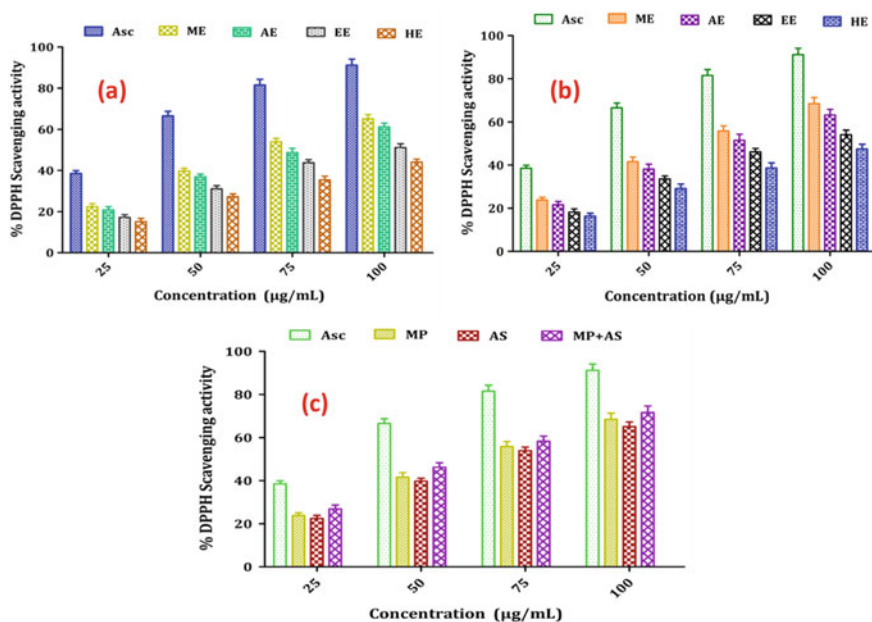
### **Assay of Cyclooxygenase**

Cyclooxygenase enzymatic assay was determined using the method reported by Copeland with slight modifications [20].

## **3 Result and Discussion**

### **3.1 *DPPH Free Radical Scavenging Assay***

By using DPPH, a stable free radical, the hexane, ethyl acetate, methanol and water extracts from *Mucuna pruriens* leaves and *Allium sativum* bulbs have been analyzed for antioxidant property. With an increase in the extract range of 25–100 µg/ml, antioxidant activity in all hexane, ethyl acetate, methanol, and water extracts was found to be increased, but less compared to standard ascorbic acid. The radical DPPH contains an odd electron that is responsible for absorption at 517 nm of and is visible in deep purple colour.



**Fig. 1** Antioxidant assay of DPPH. *Allium sativum* (a); *Mucuna pruriens* (b); and combination of *Mucuna pruriens* and *Allium sativum* (c). Asc: Ascorbic acid; ME: methanolic extract; AE: Aqueous extract; EE: Ethyl acetate extract; HE: Hexane extract; MP+AS (*M. pruriens* + *A. sativum*) (Data are expressed as the mean  $\pm$  SD of triplicate)

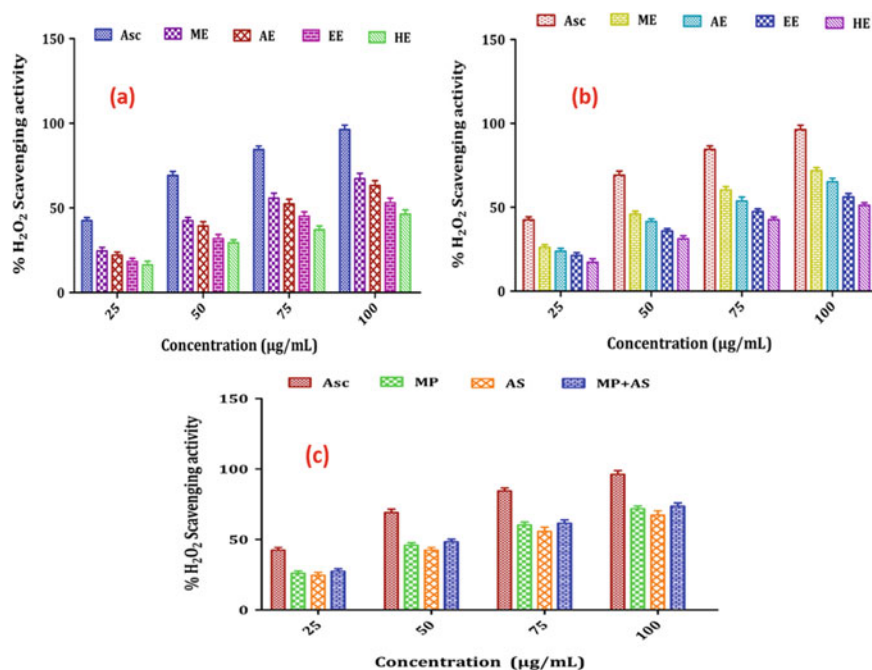
If DPPH accepts an  $e^-$  donated by an antioxidant compound, it is decolorized and can be measured quantitatively from the changes in the absorbance. The hexane showed less activity (Fig. 1a, b) and the methanol extracts showed the strongest activity in *M. pruriens* and *A. sativum*. There was also a dose-dependent increase in total antioxidant activity. The strongest activity (Fig. 1c) was shown by combination of the two plant methanol extracts, i.e., *M. pruriens* and *A. sativum* in different concentrations.

DPPH is commonly used to determine the effect of natural antioxidants on scavenging capacity of free radical. DPPH is a stable free radical at room temperature. The absorption vanishes as the electron is paired in the presence of free radical scavenging resulting in discoloration. The methanol extracts, the more polar solvent extracts, are active antioxidants compared to the non-polar hexane extract in the DPPH assay [21] based on the results obtained. The antioxidant effect of plant bioactive products is mainly due to radical scavenging of phenolic compounds such as flavonoids, tannins, polyphenols, terpenes and phenols [22].

### 3.2 Hydrogen Peroxide Scavenging Assay

The activity of the *Mucuna pruriens* leaf and *Allium sativum* bulb extracts on hydroxyl radical and the combination of *Mucuna pruriens* methanol extract and *Allium sativum* was shown in Fig. 2a–c). As the, the plant extracts showed scavenging activity against induced hydroxyl radical, from the ascorbic acid that was used as positive control. Antioxidant activity was found to be lower in all hexane, ethyl acetate, methanol, and water extracts compared to standard ascorbic acid, which is responsible for the 230 nm absorption.

There was also a dose-dependent increase in  $H_2O_2$  radical scavenging activity. The radical scavenging activity also increased with an increase in extract concentration in the range of 25–100  $\mu\text{g/ml}$ . The hexane extract showed the lowest activity and the highest activity was demonstrated by the methanol extract. Both methanolic plants extracts were mixed in equal proportion and the assay was performed. This displayed the best radical scavenging behaviour of  $H_2O_2$ . But it was slightly lower than the Ascorbic acid.



**Fig. 2**  $H_2O_2$  radical scavenging activity. *Allium sativum* (a); *Mucuna pruriens* (b); and combination of Methanol extracts of *Mucuna pruriens* + *Allium sativum* (c). Asc: Ascorbic acid; ME: methanolic extract; AE: Aqueous extract; EE: Ethyl acetate extract; HE: Hexane extract; MP + AS (*M. pruriens* + *A. sativum*) (Data are expressed as the mean  $\pm$  SD of triplicate)

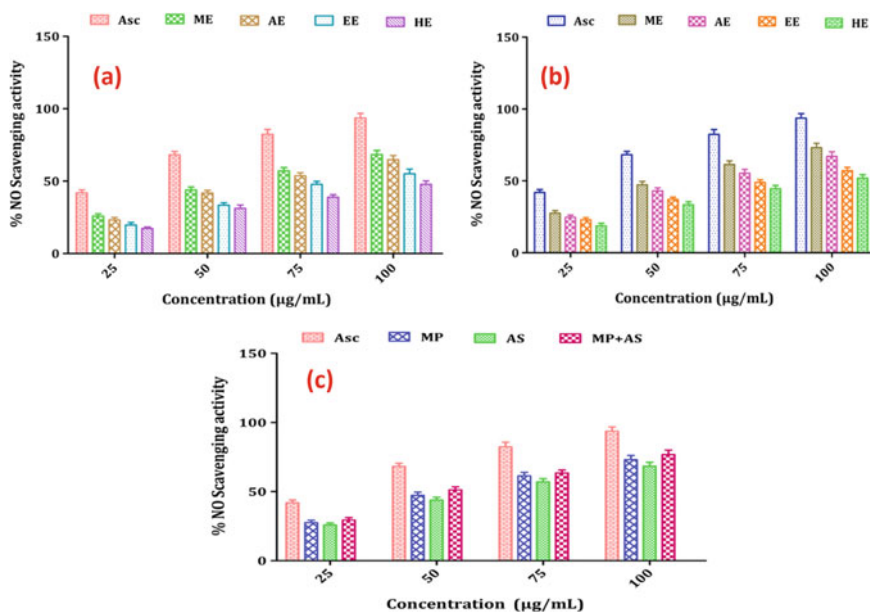
Hydroxyl radical is a highly reactive oxygen-centered radical formed by the reaction of different hydroperoxides with metal ions in transition. It attacks in membranes proteins, DNA, polyunsaturated fatty acids, and most biological molecules [23]. Hydrogen peroxide is a weak oxidizing agent that can directly inactivate a few enzymes, normally by oxidizing groups of basic thiol (–SH).

### 3.3 Nitric Oxide Radical Scavenging Assay

The nitric oxide scavenging activity of *Mucuna pruriens* leaf and *Allium sativum* bulb extract was shown in Fig. 3a, b and the combination of *Mucuna pruriens* methanol extract and *Allium sativum* in Fig. 3c.

The plant extracts showed scavenging activity against standard ascorbic acid that was used as the positive control. Antioxidant activity in all extracts of ethyl acetate, hexane, water and methanol was found to be lower compared to standard ascorbic acid, which is responsible for absorbance at 540 nm.

It was also found that NO scavenging activity increased in a dose-dependent manner. The radical scavenging activity also increased with an increase in extract concentration in the range of 25–100 µg/ml. The hexane extract displayed the least



**Fig. 3** Nitric Oxide radical scavenging assay. *Allium sativum* (a); *Mucuna pruriens* (b); and combination of *Mucuna pruriens* + *Allium sativum* (c). Asc: Ascorbic acid; ME: methanolic extract; AE: Aqueous extract; EE: Ethyl acetate extract; HE: Hexane extract; MP + AS (*M. pruriens* + *A. sativum*) (Data are expressed as the mean  $\pm$  SD of triplicate)

activity and the strong activity was shown by the methanol extract. The methanolic extracts of the both plants were combined in the same ratio and the assay was conducted. It showed the best activity of radical scavenging. But it was slightly lower than the Ascorbic acid standard.

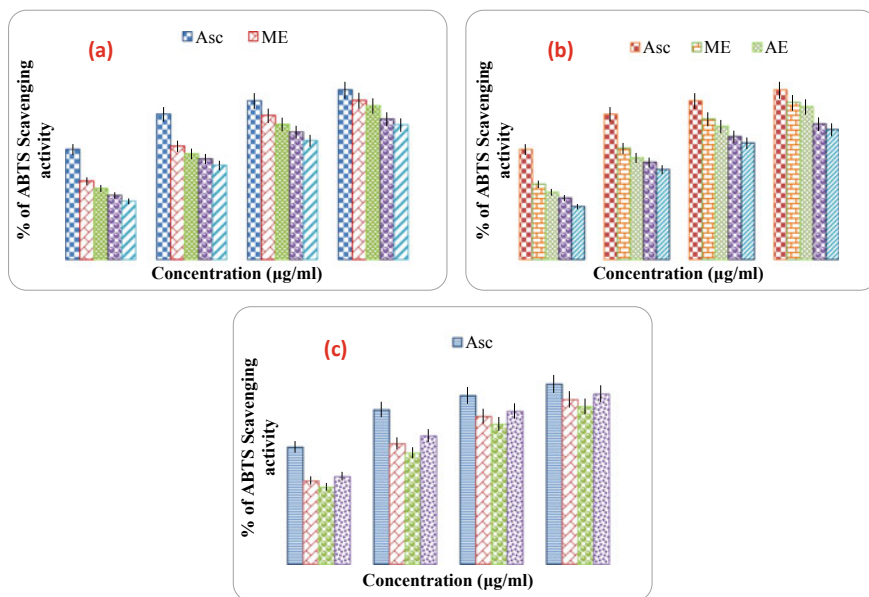
Nitric oxide is a significant chemical mediator produced by macrophages, endothelial cells, neurons and involved in regulating different physiological processes. Nitric oxide is well known to play an important role in numerous inflammatory processes such as juvenile, carcinomas, cytotoxic effects found in various disorders such as diabetes, HIV, Alzheimer's, and arthritis [24]. Nitric oxide or reactive nitrogen compounds, formed with oxygen or superoxides such as  $\text{NO}_2$ ,  $\text{N}_2\text{O}_4$ ,  $\text{N}_3\text{O}_4$ ,  $\text{NO}_3^-$ , and  $\text{NO}_2^-$  during their reaction are very reactive. Such compounds are responsible for altering most cellular components of structural and functional behaviour [25]. It has been reported that phenolic and flavonoids compounds are associated with antioxidant action in biological systems and act as singlet oxygen scavengers and free radicals [26, 27]. The scavenging operation of nitric oxide is due to the presence in plants of phenolic and flavonoids compounds [28–30].

### 3.4 ABTS Assay

The garlic bulb and *M. pruriens* leaves methanolic extracts were rapid and effective ABTS radical scavengers (Fig. 4a, b) and this activity was comparable to that of ascorbic acid. Methanol showed the best activity in comparison with hexane, ethyl acetate and water among all extracts. The combination of garlic bulb and *M. pruriens* leaves extracts of methanol also showed the best activity with increased concentrations (Fig. 4c).

The ABTS assay is based on antioxidant inhibition of radical cation absorbance,  $\text{ABTS}^+$ , which has a characteristic 734 nm wavelength. The order of the extract's radical activity of ABTS scavenging was nearly similar to that observed for DPPH. It showed powerful effects of scavenging against ABTS. The ABTS assay is a widely accepted antioxidant assay to screen the total antioxidant power of vegetables, fruits, plants and foods [17].

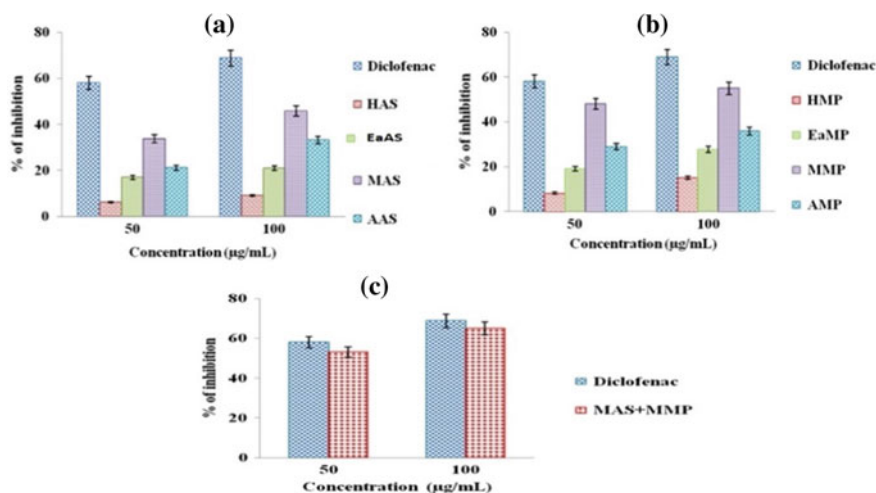
Antioxidants are important for human disease prevention. Compounds with antioxidant activity can function as free radical scavengers, pro-oxidant metal complexers, single-oxygen quenchers or reactive oxygen species and reduction agents, thus protecting the body from degenerative diseases such as cancer [31]. Many factors have been identified, such as stereo selectivity of radicals or extract solubility in different test systems, affecting the ability of extracts to react and quench different radicals [32].



**Fig. 4** ABTS radical scavenging activity. *Allium sativum* (a); *Mucuna pruriens* (b); and combination of *Mucuna pruriens* + *Allium sativum* (c). Asc: Ascorbic acid; ME: methanolic extract; AE: Aqueous extract; EE: Ethyl acetate extract; HE: Hexane extract; MP + AS: (*M. pruriens* + *A. sativum*) (Data are expressed as the mean  $\pm$  SD of triplicate)

### 3.5 Protein Denaturation

Protein denaturation is a well known cause of inflammation. The ability of different extracts of *Mucuna pruriens* (leaves) and *Allium sativum* (bulbs) to inhibit protein denaturation was analyzed in a dose-dependent manner as part of the investigation on the mechanism of anti-inflammatory activity and compared to that of Diclofenac sodium, a standard anti-inflammatory drug. They have been effective in inhibiting denaturation of heat-induced albumin. It can be confirmed from the results of this study that methanolic extracts from both plants have been active in inhibiting heat-induced denaturation of albumin. The results are displayed in Fig. 5a, b. Methanolic extracts of both plants were mixed in 1:1 ratio and tested for anti-inflammatory activity. The results showed the highest inhibition and are represented in Fig. 5c.



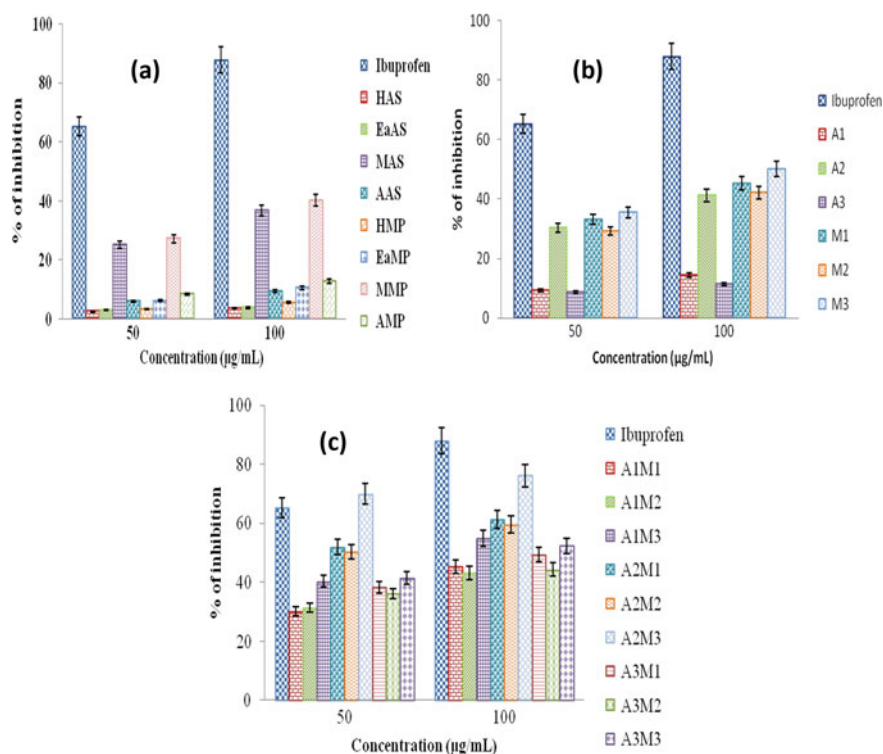
**Fig. 5** Inhibition of Protein denaturation. *Allium sativum* (A), *Mucuna pruriens* (B) and combination of *Mucuna pruriens* + *Allium sativum* methanol extracts (C). MAS: Methanolic extract of *Allium sativum*; MMP: Methanolic extract of *Mucuna pruriens*

### 3.6 Anti-inflammatory Activity

Results showed that the *M. pruriens* leaf extracts and *A. sativum* bulb extracts have good anti-inflammatory properties, but among all the extracts, the methanol extracts showed the best anti-inflammatory activity by inhibition of COX (Fig. 6a). Fractions were collected by silica gel column chromatography. For inhibition of cyclooxygenase activity (Fig. 6b), all fractions of two plant extracts were checked. Three best inhibiting fractions were collected from each extract and combined to check the potent inhibition (Fig. 6c) in all combinations in 1:1 ratio.

For further analysis, the best combination was used. These activities may be due to the strong presence of terpenoids, polyphenolic compounds and flavonoids that serve as free radical inhibitors or scavengers or may act as a primary antioxidant inhibiting inflammation.





**Fig. 6** In vitro anti-inflammatory activity by inhibition of COX enzyme by **a** plant extracts, **b** fractions and **c** combinations of fractions of *Allium sativum* (bulbs) and *Mucuna pruriens* (leaves). (HAS—Hexane extract of *Allium sativum*, EaAS—Ethyl acetate extract of *Allium sativum*, MAS—Methanol extract of *Allium sativum*, AAS—Aqueous extract of *Allium sativum*, HMP—Hexane extract of *Mucuna pruriens*, EaMP—Ethyl acetate extract of *Mucuna pruriens*, MMP—Methanol extract of *Mucuna pruriens*, AMP—Aqueous extract of *Mucuna pruriens*, A<sub>1</sub>, A<sub>2</sub>, A<sub>3</sub> are fractions of methanolic extract of *Allium sativum* collected by Column chromatography, M<sub>1</sub>, M<sub>2</sub>, M<sub>3</sub> are fractions of methanolic extract of *Mucuna pruriens* collected by Column chromatography, A<sub>1</sub>M<sub>1</sub>, A<sub>1</sub>M<sub>2</sub>, A<sub>1</sub>M<sub>3</sub>, A<sub>2</sub>M<sub>1</sub>, A<sub>2</sub>M<sub>2</sub>, A<sub>2</sub>M<sub>3</sub>, A<sub>3</sub>M<sub>1</sub>, A<sub>3</sub>M<sub>2</sub>, A<sub>3</sub>M<sub>3</sub> are combinations of the fractions of both the plant extracts in all combinations) (Data are expressed as the mean  $\pm$  SD of triplicate)

## 4 Conclusion

Protein denaturation and COX inhibition methods were used to test the in vitro anti-inflammatory property and several other qualitative, biochemical, antioxidant analyzes were also performed. The extracts of selected plant, i.e. *Allium sativum* and *Mucuna pruriens* were prepared with different solvents. Among them the hexane and methanol showed less and more antioxidant activity respectively, but the best performance was shown by the combination extract of *Allium sativum* and *Mucuna pruriens*. The above activities can be attributed to the presence of polyphenolic compounds such as alkaloids, flavonoids, tannins, etc. in the plant extracts. The

extract fractions could have been used as free radical inhibitors or scavengers or as primary oxidants that inhibits heat-induced albumin denaturation and cyclooxygenase of white blood cells. Our study revealed that *Mucuna pruriens* and *Allium sativum* possess more antioxidant and anti-inflammatory properties, the methanol extract showed the highest. But surprisingly the combinations of the two methanolic plants extracts showed the best activity than the methanolic extracts of individual plants. The higher dosage of the combination extracts has minimized side effects. This research helps to know the bioactive compounds that are responsible for these activities and to pursue new phytotherapeutics against inflammatory diseases and many more oxidative stress-related diseases.

## References

1. S. Bukke, P.S. Raghu, G. Sailaja, T.R. Kedam, The study on morphological, phytochemical and pharmacological aspects of *Rhinacanthus nasutus*. (L) Kurz (a review). *J. Appl. Pharm. Sci.* **1**(8), 26–32 (2011)
2. C. Sangita, C. Priyanka, D. Protapaditya, B. Sanjib, Evaluation of in vitro anti-inflammatory activity of coffee against the denaturation of protein. *Asian Pac. J. Trop. Biomed.* **2**, 178–180 (2012)
3. D. Arya, V. Patni, Pharmacognostic profile and phytochemical investigation of *Pluchea Lanceolata* Oliver & Hiern in vivo and in vitro. *Int. J. Pharm. Sci. Rev. Res.* **22**(2), 157–161 (2013)
4. A. Gurib-Fakim, Medicinal plants: traditions of yesterday and drugs tomorrow. *Mol. Aspects Med.* **27**, 1–93 (2006)
5. V.R. Winrow, P.G. Winyard, C.J. Morris, D.R. Blake, Free radicals in inflammation: second messengers and mediators of tissue destruction. *Br. Med. Bull.* **49**, 506–522 (1993)
6. S.R. Maxwell, Prospects for the use of antioxidant therapies. *Drugs* **49**, 345–361 (1995)
7. A. Hausladen, J.S. Stamer, Nirosoative stress method in enzymology. **300**, 389–395 (1999)
8. G. Leelaprakash, S. Mohan Dass, In vitro anti-inflammatory activity of methanol extract of *Enicostemma axillare*. *Int. J. Drug Dev. Res.* **3**, 189–196 (2010)
9. P.V. Ingle, D.M. Patel, C-reactive protein in various disease conditions—an overview. *Asian J. Pharm. Clin. Res.* **4**(1), 9–13 (2011)
10. S. Jamuna, S. Paulsamy, K. Karthika, Screening of in vitro antioxidant activity of methanolic leaf and root extracts of *Hypochoeris radicata* L. (*Asteraceae*). *J. Appl. Pharm. Sci.* **2**(7), 149–154 (2012)
11. B.J. Divya, B. Suman, M. Venkataswamy, K. ThyagaRaju, The traditional uses and pharmacological activities of *Mucuna pruriens* (L) DC: a comprehensive review. *Indo Am. J. Pharm. Res.* **7**(01), 7516–7525 (2017)
12. W. Lewis, M. Elvin-Lewis, *Medical Botany: Plants Affecting Human Health*, 2nd edn. (Wiley, New York, 2003)
13. B.J. Divya, B. Suman, L. Lakshman Kumar, M. Venkataswamy, B. Eswari, K. Thyagaraju, The role of *Allium sativum* (Garlic) in various diseases and its health benefits: a comprehensive review. *Int. J. Adv. Res.* **5**(8), 592–602 (2017)
14. M. Burits, F. Bucar, Antioxidant activity of *Nigella sativa* essential oil. *Phytother. Res.* **14**(5), 323–328 (2000)
15. M. Vijayabhaskaran, N. Venkateshwaramurthy, G. Babu, P. Perumal, In vitro antioxidant evaluation of *Pseudarthria viscid.* *Int. J. Curr. Pharm. Res.* **2**, 21–23 (2010)
16. D.C. Garratt, *The Quantitative Analysis of Drugs*, vol. 3 (Japan, Chapman and Hall Ltd, 1964), pp. 456–458

17. N. Pellegrini, A. Proteggente, M.Y. Pannala, C. Rice-Evans, Antioxidant activity applying an improved ABTS radical cation decolorization assay. *Free Radic. Biol. Med.* **26**, 1231–1237 (1999)
18. S. Sakat, A.R. Juvekar, M.N. Gambhire, In vitro antioxidant and anti-inflammatory activity of methanol extract of *Oxalis corniculata* Linn. *Int. J. Pharma Pharmacol. Sci.* **2**(1), 146–155 (2010)
19. B. Arne, Separation of white blood cells. *J. Nat.* **204**, 793–794 (1964)
20. R.A. Copeland, J.M. Williams, J. Giannaras, S. Nurnberg, M. Covington, D. Pinto, S. Pick, J.M. Trazaskos, Mechanism of selective inhibition of the inducible isoform of prostaglandin G/H synthase. *Proc. Natl. Acad. Sci. U.S.A.* **91**(23), 11202–11206 (1994)
21. Z.A. Zakaria, M.S. Rofiee, L.K. The, M.Z. Salleh, M.R. Sulaiman, M.N. Somchit, *Bauhinia purpurea* leaves extracts exhibited in vitro antiproliferative and antioxidant activities. *Afr. J. Biotech.* **10**(1), 65–74 (2011)
22. M.A.A. Rahman, S.S. Moon, Antioxidant polyphenol glycosides from the plant *Draba nemorosa*. *Bull. Korean Chem. Soc.* **28**(5), 827–831 (2007)
23. O.I. Aruoma, Free radicals, antioxidants and international nutrition. *Asia Pac. J. Clin. Nut.* **8**, 53 (1999)
24. R. Puja, V.P. Karthik, A comparative study of in-vitro nitric oxide scavenging activity of *Balofloxacinvs prulifloxacin*. *Asian J. Pharm. Clin. Res.* **10**(1), 380–382 (2017)
25. M.R. Saha, J. Rumana, M.M.I. Vhuyian, I.J. Biva, In vitro nitric oxide scavenging activity of ethanol leaf extracts of four Bangladeshi medicinal plants. *Stamford J. Pharm. Sci.* **1**(1&2), 57–62 (2008)
26. C. Rice-Evans, J. Sampson, P.M. Bramley, D.E. Holloway, Why do we expect carotenoids to be antioxidants in vivo. *Free Rad. Res.* **26**, 381–398 (1997)
27. L.V. Jorgensen, H.L. Madsen, M.K. Thomsen, L.O. Dragsted, L.H. Skibsted, Regulation of phenolic antioxidants from phenoxyl radicals: an ESR and electrochemical study of antioxidant hierarchy. *Free Rad. Res.* **30**, 207–220 (1999)
28. O.K. Kim, A. Murakami, Y. Nakamura, H. Oihigashi, Screening of edible Japanese plants for nitric oxide generation inhibitory activities in RAW 264.7 cells. *Cancer Lett.* **125**, 199–207 (1998)
29. H.K. Kim, B.S. Choen, Y.H. Kim, S.Y. Kim, H.P. Kim, Effects of naturally occurring flavonoids on nitric oxide production in the macrophage cell line RAW 264.7 and their structure activity relationship. *Biochem. Pharmacol.* **58**, 759–765 (1999)
30. G.C. Jagetia, S.K. Rao, M.S. Baliga, S.K. Babu, The evaluation of nitric oxide scavenging activity of certain herbal formulations in vitro: a preliminary study. *Phytother. Res.* **18**(7), 561–565 (2004)
31. V. Steenkamp, M.J. Stewart, L. Chimuka, E. Cukrowska, Uranium concentrations in South African herbal remedies. *Health Physiol.* **89**, 79–83 (2005)
32. L. Yu, S. Haley, J. Perret, M. Harris, J. Wilson, M. Qian, Free radical scavenging properties of wheat extracts. *J. Agric. Food Chem.* **50**, 1619–1624 (2002)

# A Novel Algorithm for Quality Evaluation Metrics of Fused Live Video Frames



K. Sai Prasad Reddy, K. Nagabhushan Raju, and D. Sailaja

**Abstract** Image fusion technique is playing important role in image processing. The main objective of image fusion is to integrate the information from several input images into a single image. The consequence of fused image consists of more precise information when compared to any of the input images. Image fusion plays pivotal role in image reconstruction. We have developed novel algorithm and implemented to evaluate quality metrics of fused video frames. This algorithm determines the quality performance between the fused video frame and unprocessed input video frame. The focal point of this paper is evaluating quality of fused video frame using structural similarity index (SSIM) and visual information fidelity (VIF) assessment methods.

**Keywords** Video frame fusion · Pixel level fusion · Simple average method · Laplacian operator · Sobel operator · Structural similarity index · Visual information fidelity

## 1 Introduction

Image and video fusion is upcoming as a essential technology which finds applications in the field of Navigation, Object recognition, Medical diagnosis, remote sensing, military applications, etc. [1]. The procedure of incorporating the appropriate details from a set of input video frames of the same scene into a single video

---

K. Sai Prasad Reddy (✉)

Department of Electronics, Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India  
e-mail: [ksaiprasadreddy@yahoo.com](mailto:ksaiprasadreddy@yahoo.com)

K. Nagabhushan Raju

Department of Instrumentation, Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India  
e-mail: [knrbhushan@yahoo.com](mailto:knrbhushan@yahoo.com)

D. Sailaja

S S B N Degree College (Autonomous), Anantapur, Andhra Pradesh, India  
e-mail: [sailajabhushan@gmail.com](mailto:sailajabhushan@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_15](https://doi.org/10.1007/978-3-030-46939-9_15)

frame can be used to give improved performance for visualization [2]. The fused video frame should protect all appropriate particulars from the input images. The video frame fusion should not establish artifacts which causes to incorrect judgment. Fusion on video frames can be performed on pixels and features. Pixel-level method can be implemented directly on each pixel of the video frame to increase the content correlated with each pixel in a video frame formed by combination of several input video frames which preserves most of the appropriate information [3]. The fused video frame can be formed either by pixel-by-pixel or by fusion of related neighborhoods of pixels in each of the video frame. The enhancement in quality by pixel-level fusion can be evaluated in video processing tasks like segmentation, video frame feature extraction In feature-level method features like edges, shape, contrast, texture and regions are extracted from source video frames and integrates all the features into a single video frame which is complete and more suitable for further video processing tasks. In this paper, we have developed algorithms for fusion between the layers of different color models in the video frame by implementing Simple Average Method. The paper is intended in discussing successful pixel level image fusion algorithm.

## 2 Edge Detection Operators

An edge detection operator finds the boundary of an image. The utilization of an edge detection mechanism is more useful in various conditions and a variety of processes have been applied from the initiation of video processing.

### 2.1 Laplacian Edge Detection Operator

This is a second order derivative operator used to evaluate edges in video frame [4–14]. Edge detection with the Laplacian enhances the contrast at edges and images emerge sharper to the watcher. In this operator there are two categories. Positive Laplacian mask and Negative Laplacian mask. Laplacian mask obtains out edges as Inward edges and outward edges [15] (Fig. 1).

**Fig. 1** Laplacian operator

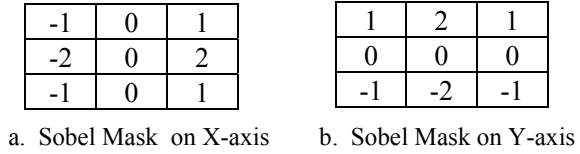
0	1	0
1	-4	1
0	1	0

a. 3X3 Positive Laplacian Mask

0	-1	0
-1	4	-1
0	-1	0

b. 3X3 Negative Laplacian Mask

**Fig. 2**  $3 \times 3$  Sobel convolution kernels



### 2.2 Sobel Edge Detection Operator

This operator is a first order derivative edge detector. This operator is composed of couple of kernels along x-axis and y-axis as shown in Fig. 2. These two kernels are meant to respond most extremely on edges which are in vertical position and horizontal position of that pixel grid [5–14]. Two kernels are applied separately on input image for gradient components which are represented by  $G_x$  and  $G_y$ . These kernels are set together in order to get complete magnitude and the orientation of that gradient [15].

Gradient magnitude is determined by using the formula [1]

$$|G| = \sqrt{(G_x^2 + G_y^2)} \tag{1}$$

An approximate magnitude is calculated using the following equation

$$|G| = |G_x| + |G_y| \tag{2}$$

Angle of orientation is represented by  $\theta$  and can be calculated by using the following equation [3]

$$\theta = \arctan\left(\frac{G_y}{G_x}\right) \tag{3}$$

### 3 Structural Similarity (SSIM) Index

Structural Similarity (SSIM) index is a technique to determine the quality of images and video frames [8–16]. SSIM is applied to evaluate the resemblance between two images. The SSIM index is a FR metric i.e. the measurement or prediction of image quality based on distortion-free image as reference. Structural Similarity (SSIM) index considers image deprivation as change in structural information [9–16]. As the pixels are spatially closer these will have strong inter-dependencies. Inter dependencies holds notable data related to object structure. Occurrence of luminance masking tends to be less visible in bright regions where as contrast masking is a occurrence where visibility of distortions are smaller extent [16]. The evaluation

between two windows and of size  $N \times N$  is

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu^2x + \mu^2y + C_1)(\sigma^2x + \sigma^2y + C_2)} \quad (4)$$

Average value of  $x$  is represented by  $\mu_x$

Average value of  $y$  is represented by  $\mu_y$

$\sigma^2x$  is denoted as the value of variance of  $x$

$\sigma^2y$  is denoted as the value of variance of  $y$

$C_1 = (K_1L)^2$ ,  $C_2 = (K_2L)^2$  are two constants

$L$  denotes pixel-values

$K_1$  and  $K_2$  are constants having default values 0.01 and 0.03 respectively.

Structural Similarity Index should satisfy the principle  $\text{SSIM}(x, y) = \text{SSIM}(y, x)$ . To evaluate the video frame quality, formula is applied on luma and chromatic values. Structural Similarity Index value ranges in between 0 and 1.

## 4 Visual Information Fidelity (VIF)

Visual Information Fidelity is a metrics which is used to evaluate the distorted image information with respect to reference image information assuming that reference image is having perfect quality [10]. Visual Information Fidelity (VIF) metrics is based on the quantity of mutual sharing of information between the reference images and distorted images. The visual quality of the distorted image totally depends upon respective information present in the distorted image. The loss of video information is called as distortion [11]. This distortion is used to compute the Video Quality Assessment metrics. The performance of Visual Information Fidelity (VIF) is far better than other existing Video Quality Assessment metrics. Computational complexity is major disadvantage of Visual Information Fidelity (VIF). Visual information fidelity measurement investigates the relationship between video information and visual quality. Visual Information Fidelity is based on the quantity of data shared by the reference and distorted images [12]. The visual quality of the distorted image is stoutly correlated to relative information existing in the distorted image. The distortion is considered as the loss of image data and is used to determine the Image Quality Assessment metrics [13]. The Visual Information Fidelity metrics have shown enhanced performance over lots of the available Full Reference Image Quality Assessment algorithms.

## 5 Video Frame Fusion

Simple Average Method: By implementing this technique the resultant fused video frame is acquired by taking the average intensity of corresponding pixels from the input video frames.

$$F(i, j) = \frac{(A(i, j) + B(i, j))}{2} \quad (8)$$

Simple Average Method fusion technique is implemented between video frame layers of RGB, XYZ, YCbCr and YUV color models. In this method Red, Green and Blue layers are considered as base layers which are used to represent video frame. The fusion is implemented between edge detected layers of Red and X, Red and Y, Red (R) and Z, Green (G) and X, Green (G) and Y, Green (G) and Z, Blue (B) and X, Blue (B) and Y and Blue (B) and Z and corresponding Structural Similarity (SSIM) Index and Visual Information fidelity (VIF) metrics are calculated. The same process is implemented to obtain the fused layers between RGB & YCbCr and RGB & YUV color models [14]. The below steps explain the proposed algorithm.

- Step 1: Construction input video
- Step 2: Selection of source for acquisition of Video frames
- Step 3: Properties analysis of video source object
- Step 4: Video frames stream preview
- Step 5: Obtaining and exhibition of video frame
- Step 6: Separation of Red, Green and Blue layers of Video frame
- Step 7:  $3 \times 3$  Sobel mask followed by Canny edge detection operator is convolved with Red component of the video frame to identify the edges and the resultant is edge detected Red layer.
- Step 8: Concatenating Red and Green layers to attain edge detected video frame. Repeat the process to obtain fused layers G & B, B & R.
- Step 9: Evaluating SSIM value for Red and Green. Repeating process to obtain SSIM & VIF values between fused layers.
- Step 10: Repeat steps 6–9 to find SSIM & VIF values for XYZ video frame, YCbCr video frame and YUV video frames.
- Step 11: Applying Laplacian mask by repeating steps 6–10.

## 6 Experimental Results

Experimental analysis is carried out by using TEN video frames to assess the performance of the edge detected algorithms based on layer fusion by implementing edge detector. The host video frame is of size  $512 \times 512$ . For the proposed fusion technique algorithm, Sobel & Laplacian masks are applied between layers of Red, Green, Blue, XYZ, YCbCr and YUV color models. The respective SSIM and VIF



are computed and shown in Table 1. Fused video frames of Red & Y, Green & U, and Blue & V are shown in from Figs. 3, 4 and 5 by applying Sobel. Fused video frames of Red & Y, Green & Cb and Blue & Cr are shown in from Figs. 6, 7 and 8 applying Laplacian operator.

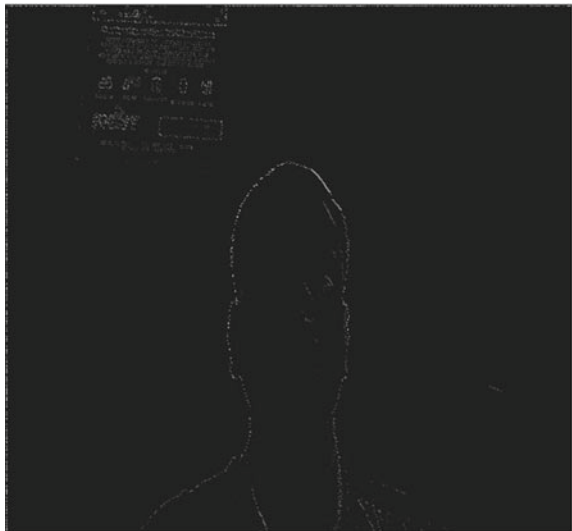
**Table 1** Performance analysis of video frame edge detection algorithm based on layer fusion between different color models

Fused video frame layers	Sobel operator		Laplacian operator	
	SSIM	VIF	SSIM	VIF
R and X	0.9926	0.8370	0.9935	0.8507
R and Y	0.9921	0.8816	0.9933	0.8473
R and Z	0.9881	0.8790	0.9936	0.8525
G and X	0.9983	0.9765	0.9936	0.8584
G and Y	0.9968	1.0241	0.9934	0.8550
G and Z	0.9920	1.0220	0.9937	0.8602
B and X	0.9980	0.9765	0.9936	0.8568
B and Y	0.9968	1.0241	0.9934	0.8550
B and Z	0.9920	1.0220	0.9937	0.8602
R and Y	0.9873	0.8620	0.9950	0.8821
R and U	0.9840	0.6558	0.9950	0.8818
R and V	0.9839	0.6558	0.9950	0.8818
G and Y	0.9915	1.0156	0.9952	0.8899
G and U	0.9930	0.7857	0.9952	0.8897
G and V	0.9903	0.7857	0.9952	0.8897
B and Y	0.9915	1.0156	0.9952	0.8899
B and U	0.9836	0.7857	0.9952	0.8897
B and V	0.9903	0.7790	0.9952	0.8899
R and Y	0.9867	0.7760	0.9943	0.8490
R and Cb	0.9856	0.7752	0.9947	0.8767
R and Cr	0.9850	0.7749	0.9946	0.8762
G and Y	0.9921	0.9115	0.9942	0.8828
G and Cb	0.9908	0.9105	0.9949	0.8846
G and Cr	0.9904	0.9101	0.9947	0.8841
B and Y	0.9921	0.9115	0.9942	0.8861
B and Cb	0.9908	0.9105	0.9949	0.8846
B and Cr	0.9904	0.9101	0.9947	0.8841

**Fig. 3** Fused Red and Y layers—Sobel operator



**Fig. 4** Fused Green and U layers—Sobel operator



## 7 Conclusions

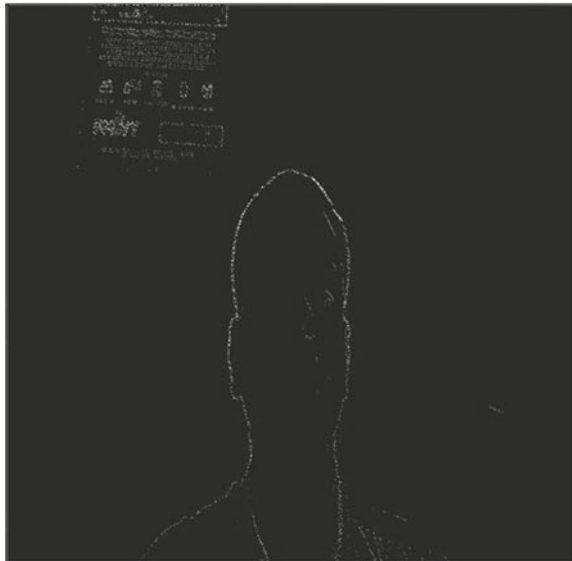
The following are the observations drawn from the proposed quality evaluation metrics of fused live video frames algorithm based on layer fusion.

- By computing SSIM values obtained from the fused video frames from each color model by applying Sobel & Laplacian operators, it is observed that Laplacian mask

**Fig. 5** Fused Blue and V layers—Sobel operator



**Fig. 6** Fused Red and Y layers—Laplacian operator



is retaining more structural information when compared to Sobel operator which indicates low data loss during transformation.

- It is observed that fused layers in YUV color model are having high values of SSIM which indicates high similarity level with Laplacian operator when compared to RGB, XYZ and YCbCr color models.

**Fig. 7** Fused Green and Cb layers—Laplacian operator



**Fig. 8** Fused Blue and Cr layers—Laplacian operator



- It is observed that Sobel operator is retaining more visual information when compared to Laplacian operator.
- It is also observed that fused layers in XYZ color model are having high value of VIF by applying Sobel operator than the rest of three color models i.e. RGB, YCbCr and YUV color models.

## References

1. H. Maitre, I. Bloch, Image fusion. *Vistas Astron.* **41**(43), 329–335 (1997)
2. Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, W. Wu, Objective assessment of multi resolution image fusion algorithms for context enhancement in night vision: a comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(1) (2012)
3. R.S. Blum, Z. Lie, *Multi-Sensor Image Fusion and Its Applications* (CRC Press, Boca Raton, 2006)
4. W. Xue, X. Mou, An image quality assessment metric based on non-shift edge, in *Proceedings of IEEE International Conference on Image Processing*, Sept 2011, pp. 3309–3312
5. H. Sponton, J. Cardelino, A review of classic edge detectors. *Image Process. Line* **5**, 90–123 (2015). <https://doi.org/10.5201/ipmap.2015.35>
6. W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Process.* **23**(2), 684–695 (2014)
7. L. Zhang, D. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**(8), 2378–2386 (2011)
8. C. Li, A.C. Bovik, Three-component weighted structural similarity index, in *Proc. SPIE*, vol. 7242 (2009), pp. 72420Q-1–72420Q-9
9. K. Gu, S. Wang, G. Zhai, S. Ma, W. Lin, Screen image quality assessment incorporating structural degradation measurement, in *Proceedings of IEEE International Symposium on Circuits System*, May 2015, pp. 125–128
10. W. Lin, C.-C.J. Kuo, Perceptual visual quality metrics: a survey. *J. Vis. Commun. Image Represent.* **22**(4), 297–312 (2011)
11. M. Unser, A. Aldroubi, M. Eden, Enlargement and reduction of digital images with minimum loss of information. *IEEE Trans. Image Process.* **4**(3), 247–257 (1995)
12. Y. Zhan, R. Zhang, A novel structural variation detection strategy for image quality assessment, in *Proceedings of IEEE International Conference on Image Processing*, Sept 2016, pp. 2072–2076
13. H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.* **14**(12), 2117–2128 (2005)
14. K. Sai Prasad Reddy, K. Nagabhushan Raju, Comparative study of Structural Similarity Index (SSIM) by using different edge detection approaches on live video frames for different color models, in *IEEE International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Apr 2018
15. K. Sai Prasad Reddy, K. Nagabhushan Raju, Video quality assessment metrics for infrared video frames using different edge detection algorithms, in *IEEE International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC-2017)*, Sept 2018
16. [https://en.wikipedia.org/wiki/Structural\\_similarity](https://en.wikipedia.org/wiki/Structural_similarity)

# Cyber Crime Investigation and Law



N. B. Chandrakala

**Abstract** Today we are living in the era of information technology where most of the activities of the private and public are dealt with the online transactions with the help of World Wide Web (www.) Anybody can access another via internet through online transaction. The entire world has become one global village. The people are connected closely with the help of internet and intranet. The internet has both uses and misuses. It is unfortunate to note that the internet is being misused by some of the criminals who use the hacking methods and blackmail the World Wide Web organisers and internet protocol users. Such criminals resort to the cyber crimes that use unauthorised access to the other's networks. Such criminal activities come under the concept of cyber crimes [1]. At this juncture the 'Cyber Law' comes to the rescue of the aggrieved. Hence, the Cyber Law was introduced in the Indian legal domain. Cyber Law deals with the crimes through internet, cyberspace and addresses the computer related legal issues and corresponding punishments. Cyber Law also addresses the sub themes such as freedom of expression, access and utilization of internet, security via online including privacy. To put it in a short way it is termed a 'the Law of the Web' [2].

**Keywords** Internet · Unauthorized access · Cyber Law · Cyber space · Hacking · Network

## 1 Introduction

Cyber Crime is treated as an unlawful act which is punishable by law. Often computer becomes either a tool or a target or both. Cyber Crime are committed are committed both in traditional way with modernity in nature. Cyber Crimes come under the purview of traditional crimes such as theft, fraud, forgery, defamation and mischief which are subject to the operation of the provisions of the Indian Penal Code and other criminal laws in India. The abuse and misuse of the internet and the computer

---

N. B. Chandrakala (✉)  
Department of Law, SPMVV, Tirupati, AP, India  
e-mail: [drkala.prof@gmail.com](mailto:drkala.prof@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_16](https://doi.org/10.1007/978-3-030-46939-9_16)

175

devices gives rise to the genesis of the new age crimes which are deftly handled by the Information Technology Act, 2000.

The Information Technology Act, 2000, the comprehensive legislation passed by the Parliament of India to address the computer related crimes and the crimes committed via internet. The Information Technology Act, 2000 also addresses the crimes in the digital domain, cyber activities including the e-commerce.

## **2 Key Points of the Information Technology Act, 2000**

- Email is recognised as a legally valid form of biz communication.
- Digital signatures have legal sanctity and validity.
- The IT Act, 2000 sanctifies and authorises the business companies to sign the certificates and related documents digitally.
- The IT Act, 2000 authorises the authorities of the Government to issue notices via internet by adopting the e-governance.
- All the communications between the companies or between the companies and Government can be made via internet.
- The IT Act, 2000 addresses the issue of cyber security and privacy of the individuals and organisation.
- The IT Act, 2000 introduced the digital platform which verifies the identity of individuals via internet.
- The IT Act, 2000 provides legal remedies for the infringement of the legal rights by the cyber criminals and redresses the harm or the loss to either individuals or companies.

## **3 Cyber Crime's Scenario in India (Few Case Studies)**

### ***3.1 The Bank NSP Case***

This case is related to the fake email id's. One of the management trainee in a bank was engaged to a marriage. The couple exchanged several emails by using the computers and internet in their company. In due course of time the couple had broken up their marriage. The bride created few fake email Id's in the name and style of 'Indian Bar Association' and sent emails to certain foreign male clients. She used the computer installed in the bank for sending her fake emails. The company in which the boy has proprietary interest sustained the loss of huge number of clients and finally the bank has become respondent in the Court of law. The bank was held liable for the fake emails sent via computers installed in the bank [3].

### ***3.2 Baze.com Case***

During the month of December 2004, the CEO of Baze.com was taken into custody for selling CD which contained certain offending content downloaded from the prohibited websites. The CDs were also sold in the open market in the Delhi city. It has been held that in case of conflict between laws, the special law(s) such as the IT Act shall prevail over general and prior laws [4].

### ***3.3 Parliament Attack Case***

This case was handled by the R&D Department of Bureau of Police, situated in Hyderabad. A laptop computer was seized from one of the terrorist who was instrumental in making an attack on the Parliament. The terrorists who were involved in the Parliament attack was gunned down on 13th December 2001 and the seized laptop computer was sent to the Computer Forensic Division of R&D Department of Bureau of Police. The seized laptop computer was found with several pieces of evidences that established the strong motive behind the terroristic attack. They fabricated the security sticker and fake Identity Cards which are used by the Ministry of Home Affairs, Central Government and affixed on their car to secure the entry into the Parliament premises. Further, the accused used emblem and seal of the Government of India for gaining entry into the Parliament. The national emblems containing three lions were scanned digitally and the Government seals were also fabricated with fake residential address of Jammu and Kashmir. After thorough investigation and careful examination it was found that all of them were forged by using computer techniques [5].

### ***3.4 Andhra Pradesh Tax Case***

The proprietor of a plastics manufacturing firm in Andhra Pradesh was taken into custody and a cash of twenty two crores rupees was recovered from him by the department of vigilance, the State of Andhra Pradesh. The authorities' asked for the proper accounts for the unaccounted cash in his possession. The suspect submitted nearly six thousand fake account vouchers which were fabricated during and after the police raids. All of them failed to prove the legitimacy of the accounts of his firm. The suspect suppresses the fact of running five other businesses under the mask of one company. In order to evade the tax, the suspect used fake account vouchers to establish the sales volume. The tax authorities exposed the dubious techniques adopted by the suspect businessman by analysing the data stored in the computers in the firm.



## **4 Technical Aspects**

The advancement in the information technology led to the new ways of criminal activities in both private and public domains [6, 7].

### ***4.1 Unauthorized Access and Hacking***

In the cyber space, the term ‘access’ means entering into, instructing or communicating with the logical, arithmetical, or memory function resources of a computer, computer system or computer network. Access refers to a type of accessing the computer via internet without the express permission from either rightful owner or the person-in-charge of a computer, system or network. Thus, hacking refers to the gaining of unauthorized access to data in a system or computer or network. Cyber hackers write or make computer programs to attack the targeted computers or network. Their motive is to destruct the software, programs of the others for illegal gains or to gain illegal satisfaction. Few of them hack the computers for personal illegal gratification. For this purpose they steal the technical information relating to the credit card, online money transfer transactions, usernames and passwords. They use such information for withdrawal of money from the others accounts. The hackers take the other’s server’s network into their control which is known as ‘web hijacking’ [8].

### ***4.2 Trojan Attack***

Trojan Attack gives an impression that is useful to the users, but in fact is damages the other’s network. Such type of fake programs are known as ‘Trojans’. Client part and Server part are conjoined in Trojans. The moment the victim is connected to the server through his computer, the hacker or attacker uses the Client to establish connection with the Server and thus start using the Trojan [9].

### ***4.3 Virus and Worm Attack***

Virus is a infected software program loaded onto a user’s computer without the user’s knowledge and performs malicious actions. It is known as ‘virus’ The malware programs that multiply like virus and infect other computers via network are known as ‘worms’ [10].

#### ***4.4 E-Mail and IRC Related Crimes***

It refers to email appears to be sent from certain source when it was actually sent from another source other than the actual source. It is known as email spoofing.

#### ***4.5 Email Spamming***

It refers to multiple users which look like a chain letters. It is known as email spamming. Attachments via emails transmitting virus, Trojans or sending a web link by which the end users download the malicious code or programs and thereby become victims [11].

#### ***4.6 Email Bombing***

It refers to the sending an identical message via email repeatedly to a particular email ID. It is known as ‘email bombing’.

#### ***4.7 IRC Related***

There are three ways to make an attack on the Internet Relay Chat. Those are Verbal attacks, Clone attacks, and flood attacks.

#### ***4.8 Denial of Service Attacks***

Flooding or overloading a computer device with more machine request more than the quantity it can handle. It leads the computer resource to crash; thereby deny access to the genuine users.

### **5 Cyber Crime Investigation**

The role of the Digitpol’s Cyber Crime and the Experts in Security Investigation is to recover and analyze the data from the internet protocol address for the forensic purpose. The global servers are monitored by the Digitpol from time to time to secure proofs by utilising the industry standard internet monitoring platforms. The

data recovered or secured from the cyber channels becomes corroborative evidence in addition to the physical substantive evidence. The Courts of law never overlook the digital evidence secured by the IT professional and considers such evidence at any stage of investigation regardless of size or source of data. The corporations, legal firms, and the agencies of the Government deploy the forensic methods digitally to encounter the cheating, fraud, tampering the financial transactions, cyber crimes, the misconduct of employees, claims of leaks etc. [12].

### ***5.1 Unauthorised Access Investigation***

The investigation and the methods of analysis of the unethical cyber fraud, hackings are monitored and the cloud storage, sky servers or physical devices cannot be used without the rightful permission of the rightful owners or occupiers. Generally, hackers gain access to the other's devices by taking advantage of their poor cyber security, malware or phishing. In case the hackers get access to your network or computers, particularly your emails, banking transactions, they change the usernames and passwords and thereby prevent the rightful user from accessing his authorised internet account or access. Impersonation by the scammers sends the messages and leads the users to click on fake web links. They may ask you to send money via internet for realizing certain financial benefits or fake financial transactions. Cyber attacks take modern shapes with sophisticated technology which makes the users to believe that their websites are genuine.

### ***5.2 Malware Analysis***

The functionality, origin, the impact of potentiality is determined by the study of malware analysis. The detect computer virus, worms, Trojan horse, root kits and backdoors. The criminals in the cyber domain use their malware or dubious software to monitor the user's online activity and thus cause irreparable damage to the computers and the storage of data. The users may often download the malware which is kept in public domain by infected email attachment and may lead the users to click a malware links via email. The usernames, passwords and other data may be stolen by using the malware forwarded to a third party. Malware comprises of virus, worms, spyware, Trojans or bots. The team which are Digitpol's specialised malware and virus analysts primarily tries to detect and remove internet threats and analyse the functions of the computer via internet. They trace the roots of the transfer of the data via web servers.

### ***5.3 Sophisticated Attacks Investigation***

The cyber criminals are so active on internet to exploit victims who are the users of computers and other internet devices. Following are the few techniques they use:

#### **Unauthorised access or hacking**

Accessing the computer or electronic device via internet or World Wide Web with the express permission of the rightful user.

#### **Malware**

It is a type of malicious software containing virus, Trojans and spyware. It causes damage to the data on computer [13].

#### **Denial of service attacks**

It is a type of cyber attack which floods a computer with unnecessary data, thereby causing overload on the computer and internet. It disables the user to make use of his computer and handle the data in proper way. The vulnerable sections for this type of attack are business organisations rather than individuals. The users are prone to such type of cyber attacks [14].

### ***5.4 DDOS: Denial of Service or Distributed Denial of Service Attacks Investigation***

Denial of service attacks are common in cyber attacks. They dump big data and flood the computers and the websites. Thus causes overload on the systems which may lead to the malfunctioning of the computer network. It disables effective working of the computer system unlike hacking or malware attack. It denies the service attack which is the result of the distributed denial of service, often a network of computers and systems.

### ***5.5 Email Fraud Investigation***

Crimes pertaining to the email frauds, email spear phishing attacks, online scams and fraud are constantly monitored by the Digitpol's Cyber and Fraud Team which are certified examiners. The task of the Digitpol is to investigate into the issue of hacking. They determine the root causes of hacking, report their findings. They are empowered to prevent the hackers from interfering with the user's network and further prevent them from further malware attacks.

### ***5.6 Phishing Attack Investigation***

Cyber criminals resort to the phishing attacks, frauds via emails, online scams, fraud through email servicers particularly the small and medium business enterprises operating in Asian countries [15]. They try to gain access via emails and steal the business information such as outstanding bill amounts, pending invoices, financial transactions and data between the supplier, vendor and the buyers. The cyber criminal find out the due invoices, send fake emails by hacking email account and make the users to pay into their own accounts. The fake emails appears to be sent from the rightful email users. There is every chance to believe that the same has been sent by the rightful users. The cyber criminal use the nominated account in the same name and style of the rightful business enterprises with or without a slight change in the nomenclature of the business enterprise which makes the end user to believe that it has been sent by the rightful user. The rightful user may believe that the bank account is in the cavity as the victim or client [16].

### ***5.7 Office 365 Phishing Attack Investigation***

The examiners duly certified by the authorities concerned assist the rightful users to prevent the phishing attacks, emails frauds and scams. They are known as Digitpol's Cyber and Fraud Team. In case of phishing attacks such as internet fraud, CEO fraud or scams the cyber crime investigation and the law come to the rescue thereby by provides cyber security. The user should act vigilantly and instantaneously with proof of fraudulent activity via cyber domain. The timely held is provided by the Digitpol to prevent or stop the transfer of funds and direct the user to report to the commission of crime to the local police.

## **6 Cyber Warfare**

The technology which is used to make an attack against nation or the national interest causing equal to the actual warfare is called 'Cyber Warfare'. It may not be equivalent term as 'war' but protects the national interest. The team works under the Digitpol's Cyber Intelligence constantly monitors the serious threats and the rogue activities committed via internet domain [17].

## 7 Conclusion

In the recent times, there is a high rise of new techniques that paved way towards the commission of cyber crimes. The cyber crime over the internet domains is one of the main threats to the mankind which has to be addressed fastidiously. There is immense need to wage a war against the cyber crime. It has to be considered as an priority to protect and promote the social, cultural and security of a country. Cyber crimes have no geographical boundaries. They are operations over cross country borders. They give rise to the technical and legal complexities particularly the cyber crime investigation and criminal prosecution to bring the cyber criminal to the book. A coordinated and comprehensive action is needed to address the problems of the cyber crimes committed via internet. The very purpose of this article is to bring awareness and spread the technical and legal aspects of the cyber crimes to those who are computer literates in general and computer illiterates in particular [18]. The victims of cyber attacks cannot keep silent and the cyber crimes cannot go unreported. Everyone of us should shoulder the responsibility to come forward to register the cyber crimes that they face in their personal and official domain to the nearest cyber police station. Unless the cyber criminal are punished, the cyber crimes will continue to rise in its own way thus cause damage to a great extent. At this juncture, one must have awareness about the cyber crimes and also should have gushing concern to prevent and stop the commission of such heinous crime.

## References

1. Aadhaar number and the name is related to IP address, date and time of authentication, device ID and its unique ID of authentication device which can be used to locate the individual
2. OECD, Guidelines Governing the Protection of Privacy and Transborder Data Flows of Personal Data, Paris (1981)
3. [http://www.indiancybersecurity.com/case\\_studies/the\\_bank\\_nsp\\_case.html](http://www.indiancybersecurity.com/case_studies/the_bank_nsp_case.html)
4. 008(105)DRJ 721 MANU/DE/0851/2008
5. AIR 2005 SC 3820
6. Sections 121, 124A, 125, 153A and 505 of the Indian Penal Code and Section 13(1)(b), 18B, 39 of Unlawful Activities (Prevention) Act, and Section 66F of Information Technology, Act
7. Google India Private Limited in O.S.No.143 of 2010
8. Sections 65, 66, 66-C and 85 of the IT Act
9. Wasi Udain Ahmed v. District Magistrate, Aligarah, 1981 CriLJ 1825
10. Anvar P.V. v. P.K. Basheer and others, (2014) 10 SCC 473
11. Section 67 of IT Act
12. Rule 11 of The Information Technology (Qualification and Experience of Adjudicating Officers and Manner of Holding Enquiry) Rules, 2003
13. Website blocking to apply existing and new legislation to a range of legitimate public policy goals that involve the Internet
14. From the modus operandi of the transaction it is clear that the transaction effects through hacking and it amounts to cyber crime
15. State of West Bengal & Ors. v. Committee for Protection of Democratic Rights, West Bengal & ors. (2010) 3 SCC 571

16. 'Phishing' is a form of internet fraud. In a case of 'Phishing', a person pretending to be a legitimate association such as a bank or an insurance company in order to extract personal data from a user such as access codes, passwords etc. which are then used to his own advantage, misrepresents on the identity of the legitimate party. Typically 'Phishing' scams involve persons who pretend to represent online banks and siphon cash from e-banking accounts after conning consumers into handing over confidential banking details
17. Section 3 and 9 in the Information Technology (Intermediaries Guidelines) Rules, 2011, the intermediary shall report cyber security incidents and also share cyber security incidents related information with the Indian Computer Emergency Response Team
18. Unlike the political fiction of a 'State', generally having geographical boundaries, a Government is a dispensation which runs the bureaucratic administration of the State at a particular point of time and cannot be identified with the State itself

# Real Time Recognition of Rashdriving and Alcohol Detection to Avoid Accidents and Drunken Driving



S. Swarnalatha, T. Srilakshmi, and K. Thilak Kumar

**Abstract** In Today's life most number of accidents are occurring because of rash driving and driving the vehicle by consuming vehicle. These are the two important reasons for occurring accidents today and may also be in future generations. So we have to take some important decisions to prevent the accidents happening with this two reasons. Now a days as technology got improved more and more we have different systems available to detect the alcohol content in drivers breath and detection of the vehicle speed which exceeds normal speed limit that pose danger to driver. More number of accidents is occurring and is increasing day by day and among these accidents more than 50% are occurring due to alcohol consumption and 50% is occurring due to rash driving. So we have to take immediate action to prevent accidents due to alcohol consumption and rash driving. To overcome all these, this paper provides a smart system that detects the drunk and driving as well as over speeding on roads using vehicular networks tech. The main objective is to stop the drunken person by traffic personnel as early as possible and save lives before the accident even happens.

**Keywords** Embedded · Drunk and drive · Rash driving · MEMS sensor · MQ3Sensor · LCD

## 1 Introduction

At present vehicle transport system takes a major role for going from one place to another place. But accidents are occurring more and more in today's life. So to

---

S. Swarnalatha · T. Srilakshmi (✉) · K. Thilak Kumar  
Department of Electronics and Communication Engineering, S.V. University College of Engineering, Tirupati, A.P., India  
e-mail: [srilakshmitelegam.15@gmail.com](mailto:srilakshmitelegam.15@gmail.com)

S. Swarnalatha  
e-mail: [swarnasvu09@gmail.com](mailto:swarnasvu09@gmail.com)

K. Thilak Kumar  
e-mail: [krsna.thilak@gmail.com](mailto:krsna.thilak@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_17](https://doi.org/10.1007/978-3-030-46939-9_17)



prevent these types of accidents drivers should be aware of all traffic rules such as traffic signals, not consuming any alcohol while driving and not to exceed the speed limit which is beyond the normal controllable speed. As the drunk and drive reduces the human perception and to recognize the vehicle coming in front and cannot even control the vehicle which causes severe life threatening accidents on roads [1]. Drivers should drive the vehicle with the concentration because all the lives of the passengers and pedestrians will be in danger if he did any mistake. So we have to develop the technology to avoid the accidents based on the alcohol consumption and rash driving. So that the technology will send the signal to the nearest police station and it also track the Coordinates of the position. Today drink driving and rash driving was one of the major causes for fatal accidents all around the world [1–4].

In South Africa (SA), drunk driving becomes ultimate causes of traffic accidents that is under serious concern. We have to reduce the accidents by giving alerts to drivers to wear the seat belt, not to consume alcohol while driving and to keep the speed limit. So necessary precautions has to be taken in the developing countries as well as developed countries. This paper will introduce the technology to avoid the accidents based on the rash driving and alcohol detection.

Many studies are performed in recent years in the viewpoint of drivers to monitor and to prevent the drunk driving. Moreover many systems are implemented, developed and used with different kinds of technology as in [5]. Some of the studies focus on preventing drivers fatigue and some other focus on real time driving pattern recognition. We consider these approaches in developed countries and not only in developing countries. From the previous studies most of the accidents are occurring due to alcohol detection and rash driving. The objective of this paper is to diminish the accidental rate and to provide safety while driving on the road. The approach implements vehicular ad hoc networks (VANETs) [6] and Internet of Things (IoT) [7, 8] technologies.

## 2 Earlier Work

Several works are carried out in the effort to reduce the accidents that often occur on the roads due to rash driving and driving when drunk. In existing method using an AlcoKey that automatically collects the driver's breath sample and triggers the vehicular control unit and that checks the content of alcohol in the driver's body and sends the signal whether the vehicle should start or stop. In method we are using alcokey product at the place of steering which checks the alcohol level in the driver by using breath samples and if the alcohol content is more than the threshold then the car locks and the engine stops all by itself.

### 3 Hardware Description

(a) **Mems sensors**

Micro-Electro-Mechanical Systems or MEMS technology is a mini sized mechanical and electromechanical elements (i.e., devices and structures) which are developed by the help of micro fabrication techniques. The physical dimension of MEMS sensor can be in the range of several millimeters to less than one micrometer. The dimensions are smaller than the width of a human hair. This sensor is a power efficient, small and full 3-axis accelerometer with signal conditioned voltage values as output. The product measuring full-scale range is of  $\pm 3$  g acceleration.

(b) **MQ3Sensor**

MQ3 alcohol detection sensor uses semiconductor with low level sensor with low cost that will find the any trace of alcohol gases in the driver body based on the collection of breath samples which can provide alcohol level present in the blood content. The sensitive material here is SnO<sub>2</sub>, whose conductivity is low in clean air. If the level of concentration of the alcohol gases increases, conductivity also increases. This MQ3 module gives both digital and analog output values. Interfacing the module with microcontrollers, Arduino boards and Raspberry pi etc. is very easy.

(c) **Atmega328**

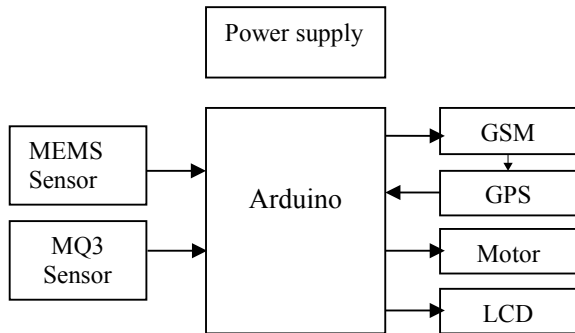
ATmega328 belongs to AVR family microcontroller and has 28 pins. It supports data up to 8 bits. It has 32 KB internal built-in memory. It has 3 built-in timers two are 8-bit and one is 16-bit timer. It dissipates low power, cost effective and has programming lock for security purposes.

ATmega328 has an adc converter which is built-in converts analog voltage into digital form (Fig. 1).

(d) **LCD Display**

LCD is used for displaying text coming from the Bluetooth module. We can display numbers, letters and graphics. LCD has 8-bit data line where it can be made to work

**Fig. 1** The connection between power supply and arduino



in 4-bit mode or 8-bit mode. Three control pins are there RS, EN and RW. To write RW pin should be low and to read it should be high. Using commands we can make letters to change their displaying position.

## 4 Methodology

By this project one knows that the rash driving or the alcohol concentration in driver were detected if rash driving is detected then this module sends the information to the nearest police station with the GPS coordinates. System utilizes MEMS and MQ3 sensors for the readings. MEMS provide the value of X, Y, Z co-ordinates as per the movement of the vehicle. According to the readings the speed of motor can be changed. If driver going very fast then it sends SMS to the control room with coordinates. If MQ3 sensor gets activated then motor stops immediately and sends SMS with GPS coordinates. If alcohol Sensor and MEMS are detected the motor will be stopped.

In this project, detecting the speed and finding whether the driver consumed alcohol or not is the main motto. The inputs from GPS, MEMS and MQ3 sensors connected to Arduino. If the MQ3 sensor is activated, immediately motor stops and sms will be sent with the coordinates. If the driver giving full acceleration through MEMS that means the motor is running very fast so called rash driving, at that time the system sends the SMS with rash driving message with GPS coordinates in it.

## 5 Result

The Vehicle speed can be tracked simply by using MEMS sensor and if the driver consumed alcohol or if the speed of the vehicle exceeds then the motor stops and the module will send the SMS to the friends and family by using GPS and GSM so that it will be safe to both driver and pedestrians on the road. Here Alcohol was detected by using MQ3 Sensor which automatically triggers the engine (motor in the module) to stop and cannot start in Figs. 2, 3, 4, 5, 6, 7, 8 and 9.

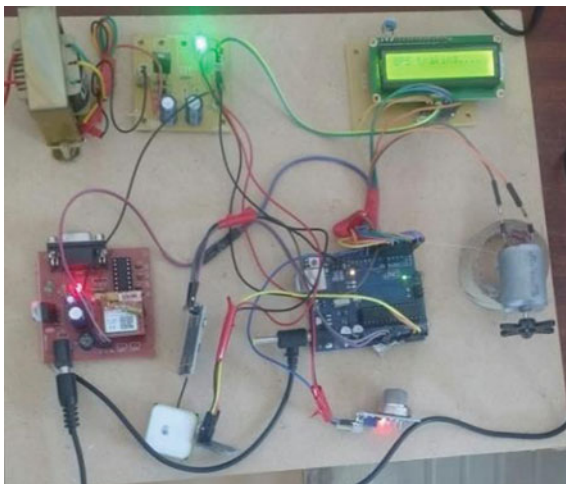
## 6 Conclusion

In this paper, we present the most efficient and effective GSM based rash driving and alcohol detection system. The MEMS sensor detects the vehicle speed and alcohol sensor will be detecting alcohol level in the blood by using breath sample and collect data and send the information to the concerned number. The main aim of this paper is to provide safety to the driver and to prevent the accidents on the road by identifying the abnormal behaviors of the driver and intimating them to the concerned persons of

**Fig. 2** Normal speed

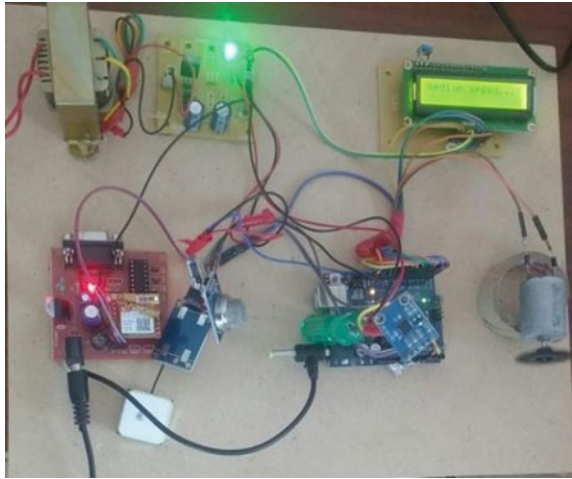


**Fig. 3** GPS tracking

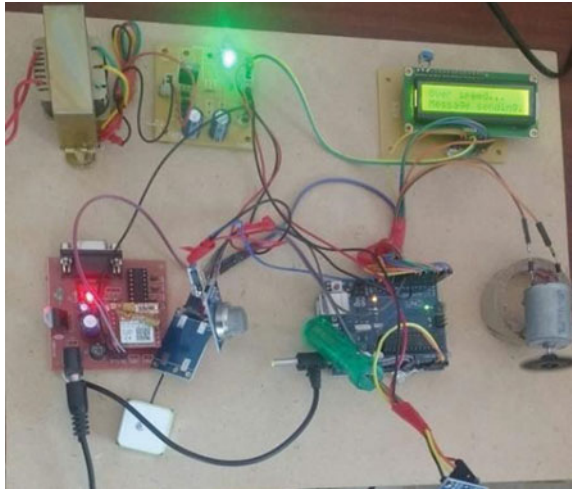


driver. During this paper, we propose a system that can effectively estimate and detect by observing some specific kinds of abnormal driving behaviors and by sensing the vehicle's speed as well as alcohol content thereby providing safety measures.

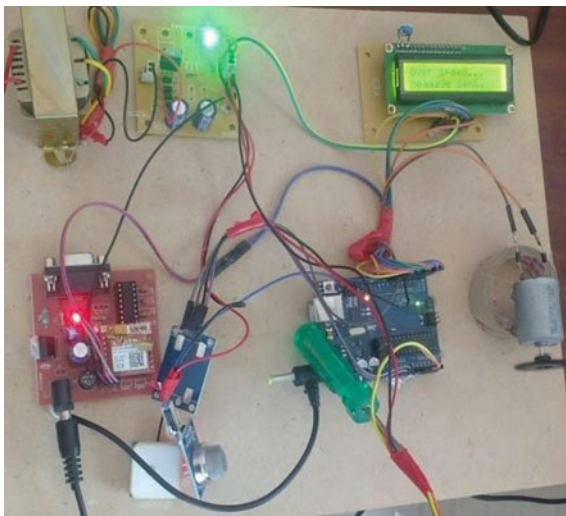
**Fig. 4** Medium speed



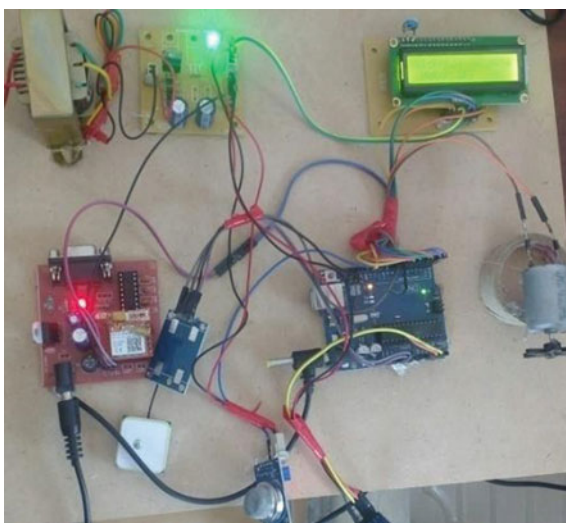
**Fig. 5** Over speed message sending



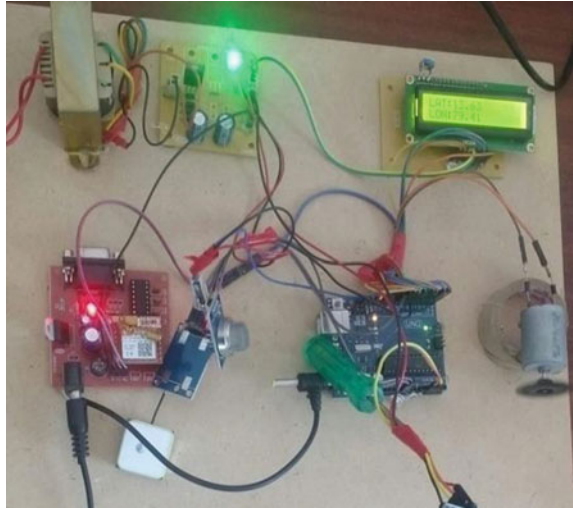
**Fig. 6** Over speed message sent



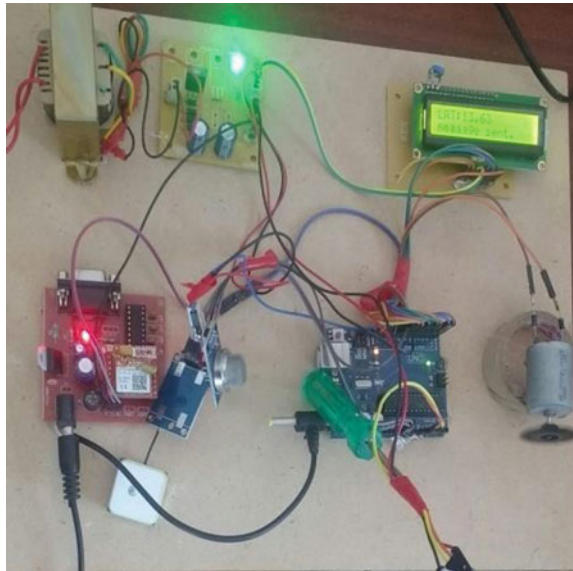
**Fig. 7** Alcohol detected motor off



**Fig. 8** Detection of latitude and longitude using GPS



**Fig. 9** Message Sent by using latitude and longitude values



## References

1. European Transport Safety Council, *Drinking and Driving in Commercial Transport* (E TSC, Brussels, 2012). [http://etsc.eu/wpcontent/uploads/Drink\\_Driving\\_in\\_Commercial\\_Transport.pdf](http://etsc.eu/wpcontent/uploads/Drink_Driving_in_Commercial_Transport.pdf). Accessed 15 Jan 2017
2. WHO Global Status Report (2015). <http://www.sadd.org.za/education/statistics?showall=&start=2>. Date Accessed 22 Aug 2016



3. Statistics. <http://www.sadd.org.za/education/statistics?showall=&start=1>. Date Accessed 22 Aug 2016
4. Y.A. Phanama, C. Duthoit, R.F. Sari, Aware-D: voice recognition-based driving awareness detection, in *22nd Asia-Pacific Conference on Communications (APCC)*, Yogyakarta (2016), pp. 90–95
5. J. Dai, J. Teng, X. Bai, Z. Shen, D. Xuan, Mobile phone based drunk driving detection, in *4th International Conference on Pervasive Computing Technologies for Healthcare*, Munich (2010), pp. 1–8
6. K.S. Xu, P. Guo, B. Xu, H. Zhou, QoS evaluation Of VANET routing protocols. *J. Netw.* **8**, 132–139 (2013)
7. A. Botta et al., Integration of cloud computing and Internet of Things: a survey. *Future Gener. Comput. Syst.* **56**, 684–700 (2016)
8. J. Gubbi et al., Internet of Things (IoT): a vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **29**(7), 1645–1660



# XGBoost Classifier to Extract Asset Mapping Features



K. Sree Divya, P. Bhargavi, and S. Jyothi

**Abstract** In steep growth in the consumption of Internet, Big Data came into picture for handling enormous amount of data. However, the data that is generated through internet has high dimensional data. So, feature engineering will be performed—to extract the best feature subset from high dimensional data. Assets are the ones to keep, expand upon, and support for the one who and what is to come. Asset mapping is a positive and charming way to learn about the community. It empowers us to contemplate where individuals live and work. It also challenges us to recognize how other people see the same community. In this paper, a model is introduced to find the required assets based on the population in the area and whether the available assets are tangible are not, is identified by extracting the features from the data gathered from the government of Andhra Pradesh. The data is pre-processed by extracting the best features in it by using feature engineering methods and classifiers like XGBoost, Random Forest and ExtraTreeClassifier. The experimental results proves that XGBoost provides the most accurate results for the specified target.

**Keywords** Feature extraction · Asset mapping · ExtraTreeClassifier · XGBoost · Random Forest classifier

## 1 Introduction

The procedure of selecting features is to choose relevant attributes for large dimensional dataset. This process involves identifying the relevant instances and removing irrelevant and redundant instances from the dataset. Using this process, can easily

---

K. Sree Divya (✉) · P. Bhargavi · S. Jyothi  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [divya.kpn@gmail.com](mailto:divya.kpn@gmail.com)

P. Bhargavi  
e-mail: [pbhargavi18@yahoo.co.in](mailto:pbhargavi18@yahoo.co.in)

S. Jyothi  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_18](https://doi.org/10.1007/978-3-030-46939-9_18)

train the data by using learning models which help for easy analysis of future prediction [1]. There are different reasons for feature selection -faster training of our data, reduces the entanglement of a model, tweak the exactness of the model and reduces overfitting. There are different models for selecting appropriate features [2, 3]. Feature engineering algorithms are classified into different types (1) Filter (2) Wrapping (3) Embedded (4) Hybrid and (5) Ensemble.

Filter takes just a part of the pertinent features and uses as a pre-handling step. This process is done using correlation matrix by using Pearson correlation for continuous instances or and Chi-Square test for categorical instances. The selection of features does not depend on any learning mechanisms. Instances are chosen based on their raw outcome in various statistical trail for the correlation with the target variable. The correlation coefficient has values between  $-1$  and  $1$ . A worth more like  $0$  suggests more fragile correlation (precise  $0$  inferring no correlation). A worth more like  $1$  infers more grounded positive correlation. A worth closer to  $-1$  suggests more grounded negative relationship.

A wrapper technique utilizes machine learning algorithm and makes its exhibition as assessment criteria. It encourages the instances to the preferred Machine Learning models and dependent on the exhibition of the model by include/evacuate the instances. This is an iterative and computationally costly procedure yet it is more precise than the filtering technique. There are distinctive wrapper strategies, for example, Backward Elimination, Forward Selection, Bidirectional Elimination and Recursive Feature Elimination.

Embedded techniques consolidate the characteristics of wrapper and filter strategies. It is executed with models that have their own worked in selecting of features strategies. The most preferable techniques are LASSO and RIDGE regression which have built-in punishment capacities by diminish over fitting. LASSO regression performs L1 regularization, gives chastisement equal to unflawed estimation of the magnitude of coactive. RIDGE regression performs L2 regularization gives chastisement equal to double of the size of coactive.

Hybrid feature selection procedure utilizes a combination of test domain filtering and resampling to refine the test domain and two instance subset assessment strategies (channel and wrapper) to choose reliable features [4]. This strategy uses both feature space and test domain in two stages. The main stage filters and resamples the test area while the subsequent stage receives a hybrid strategy by data gain, wrapper subset assessment and hereditary pursuit to locate the optimal feature space. It takes favourable circumstances of wrapper subset assessment with a lower cost and improves the exhibition of a grouping of classifiers.

Ensemble method is a learning strategy that combines several base models to improve the overall performance and produces an optimal predictive model by using Machine learning. Ensemble techniques are (1) Bootstrap aggregation (or) Bagging (decrease variance), and (2) Boosting (decrease bias) [5].

**Bagging (Bootstrap AGGREGATING)** First, make random samples of the training data set with substitution (part of training data set). At that point, construct a model

(classifier or Decision tree) for each sample. At long last, consequences of these various models are joined by utilizing average or majority voting.

**Boosting** It is an iterative method which modifies the weight of a perception dependent on the last classification. In the event that a perception was classified wrongly, it attempts to build the weight of this perception and the other way around [6].

### *1.1 Reasons Leads to Problem Definition*

Various types of feature selection method are available for different kinds of problems with like, redundancy, cost and inconsistent prediction accuracy with filter, wrapper and embedded types of feature selection procedures. To overcome this, ensemble selection of instance methods are preferred for minimal subset of relevant and non-repetitive features which generates accuracy while classification.

Asset mapping is a power-based strategic plan for community development. There are three types of asset maps: First one is Discrete plan: find the individual with new connection to lead the changes within the area (e.g. elected officials, school principals, police commissioners, pastors). Second one is Citizen Association maps: Resident Association maps: Identify stages for urban commitment and supporters in the community (for example housing colonies, grouping of youngsters, clubs). Third one is Establishment maps: Identify common organizations inside the community (for example schools, libraries, emergency clinics). The goal of asset mapping is to archive the area's current assets, consolidating these qualities into community improvement work. Resource mapping attempts to recognize the assets that are available in the area, and spotlight on the critical thinking capacities of the local's occupants. This can be practiced through community investigation, internet surfing, reviewing, and so on.

Asset maps look like a record of influential people, affiliations, and foundations inside a cluster and a data set containing academy, assets, and community linkage. Resource mapping is unique in relation to different strategies for community improvement since it focuses around what an area has rather than on what it needs, expecting that numerous answers for a community issues as of now exist in the community. In this proposed work, mainly focusing on both neighbourhood and needs in the area, which strengthen the community in all aspects.

To notify the needs and neighbourhood in a community dataset for extracting the relevant features to strengthen the community.

## **2 Data Modelling**

In asset mapping, finding institution maps in a particular area to strengthen the community and the assets are tangible for the area by fulfilling the necessities of the

people in the community range. For finding necessary institution maps, extract the features of the area.

### Proximities

When there is a large data set and it doesn't fit a  $K \times K$  lattice into fast memory, which diminishes the necessary memory size to  $K \times M$  where  $M$  is the weight of trees in forest. To accelerate the reckoning-rigorous  $s$  measure and recursive missing worth substitution, the client is provided with the choice of holding just the biggest presence to each case.

### Scaling

The propinquity between cases  $p$  and  $q$  form a matrix  $\{\text{prox}(p, q)\}$ . From the rendition, it is easy to display that this matrix is symmetric, positive definite and bounded above by 1, with the diagonal elements equal to 1. It follows that the values  $1 - \text{prox}(p, q)$  are doubled distances in a Euclidean space of dimension not greater than the number of cases [3].

### Feature selection

Feature Culling is an approach by opting the most pertinent instances from the dataset and further using the exact machine learning models for the better pursuance of the model. Huge number of inappropriate instances prolongs the training time aggressively and extends the risk of over fitting. To solve this problem, select the most pertinent feature from the high dimensional data [7].

There are multiple methods to select features from the dataset (1) Correlation (2) Univariate Statistical Tests (3) Recursive Feature elimination (4) Recursive Feature Elimination with cross-validation and (5) Boruta.

Statistical tests can be used to choose those features that have the substantial relationships with the target variable. So, method used for univariate statistical test  $\chi_c^2$  test.

Chi-square test  $\chi_c^2$  is worn for unconditional instances in a dataset. Calculating  $\chi_c^2$  for the every instance including target. Choose the required number of instances with best  $\chi_c^2$  scores. It resolves the relation between two unconditional variables of the sample and it would through back their real relationship in the population.  $\chi_c^2$  value is given by

$$\chi_c^2 = \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}} \quad (1)$$

Points Observed frequency = Total amount of observation in a class and Expected frequency = Expected number of resultants in a class, if there was no relationship between the instances and the target.

As explained in the introduction part, Chi-square test is filter based feature selection model. It exquisite if the attributes are completely independent, and is not complicated, and all instances are always related.

The embedded instances procedure are selected during training the model. The most preferable technique is LASSO, which is a linear regression with a summarization of regularization term. The results of coefficients are drawn to zero for statistically less important variables in objective function minimization [8].

Recursive Feature Elimination (RFE) is a wrapper-based election model, that eradicate the features repetitively. The model begins with the complete regression model and it has all instances by excluding the least useful diviner in each iteration. The process of the model:

Denote  $\hat{f}^n$  as a model consisting of all the instances. For every k instances,  $k = n, n - 1, \dots, 1$  reject the target with the least preferable instance, fit a new model  $\hat{f}^{n-1}$  and compute a raw score such as AIC, BIC, cross-validated  $L^2$  for regression problem or cross-validated accuracy for classification problem. Select the exact model from  $\hat{f}^n, \hat{f}^{n-1}, \hat{f}^0$  based on the computational score values.

Another and important feature selection method is Boruta. It works as a wrapper selection algorithm and provides importance for several variables in the dataset and it gives good prediction accuracy.

### 3 Methodology

Ensemble selection methods prefers requited facts to select a minimal features from a part of the original dataset. Requited Facts  $RF(X; Y)$ , is the skepticism in X due to Y. Requited Facts is defined as:

$$R(X; Y) = \sum_{x,y} g(x, y) \log_2 \frac{g(x, y)}{g(x)g(y)} \tag{2}$$

where  $g(x, y)$  is the distribution function of X and Y features, and  $g(x)$  and  $g(y)$  are probability of marginal. Requited Facts, which swiftly gain different types of consortium and generates functional relationship between features, and it explores this relationship between different pair of features in large volume of data, i.e., big data.

Consider a data set  $D_s$  with finite elements and ordered pairs  $\langle R, S \rangle$ , the no. of elements is  $n$ , and the  $r, s$  plane is isolated into little lattices. This division is called  $r \times s$  grid  $G$

$$I^*(R, S, D, i, j) = \max(R, S, D / G, i, j) \tag{3}$$

Maximum requited facts between  $r$  and  $s$  is represented in Eq. (3) where  $G$  is divided into  $i \times j$  grid.

### 3.1 Random Forest Classifier Principle

Consider a constant ensemble of classifiers  $r = r_1(x), r_2(x), \dots, r_k(x)$  and a vector  $(V, u)$  with random data. If there are any one of its outcome for a classifier  $r_k(V)$  in the ensemble, define

$$\hat{P}(A) = \text{proportion of classifier } r_k(1 \leq k \leq K) \text{ for which events } K \text{ occur.}$$

$$= \text{Empirical probability of } A \tag{4}$$

Empirical Marginal Function is

$$\hat{m}(V, u) = \left( \hat{P}_k(r_k(x) = u) - \max_{j \neq y} \left( \hat{P}_k(r_k(x) = j) \right) \right) \tag{5}$$

Exceeds to the one where the mean number of votes for a perfect class exceeds the mean number of votes for the next-perfect class.

In General, a random forest is a visionary, where it possess a collection of random base regression trees  $\{\bar{r}_n(x, \Theta, D_n), m \geq 1\}$ , where  $\Theta_1, \Theta_2, \dots$  are i.i.d. outputs of a randomizing variable  $\Theta$ . These random trees are coagulated to create an aggregated regression estimate [9]

$$\bar{r}_n(N, D_n) = E_{\Theta}[r_n(N, \Theta, D_n)] \tag{6}$$

where  $E_{\Theta}$  denotes random parameter with expectation, conditionally on  $N$  and the data set  $D_n$ .

Imagine, every separate random tree is built by all vertices of the tree are related with block cells to such an extent that at each every progression for construction of the tree, the assortment of cells related with the leaves of tree shapes a bifurcation of  $[0, 1]$  d. The root of the tree is  $[0, 1]$  d itself.

The upcoming method is then rolled for  $\lceil \log_2 k_n \rceil$  times, and  $\log_2$  is the base-2 logarithm,  $\lceil \cdot \rceil$  the ceiling function and  $k_n \geq 2$  a deterministic attribute, steadfast beforehand by the user, and possibly depends on  $n$ .

1. Every node has a coordinate of  $V = (V(1), \dots, V(d))$  is opted, with the  $j$ th feature having a probability on  $j \in (0,1)$  of being selected.
2. Each node, first the coordinate is chosen, then, the split is at the centroid for the chosen side. Each randomized tree  $r_n(V, \Theta)$  targets the mean of over all  $Y_i$  for the relating vectors  $X_i$  fall in the common cell of the random partition as  $V$  which we specified. Taking finally expectation with respect to the parameter  $\Theta$ , the random forests regression estimate takes the form

$$\bar{r}_n(V) = E_{\Theta}[rn(V, \Theta)] = E_{\Theta} \left[ \frac{\sum_{i=1}^n Y_i 1_{[V_i \in A_n(V, \Theta)]}}{\sum_{i=1}^n 1_{[V_i \in A_n(V, \Theta)]}} \right] \tag{7}$$

In selecting the best feature from the given dataset, to amend the prediction efficiency in defining new trees, Random Forest provides the best score for each feature, shows importance of each feature to train the model, and generates a new tree by maximize the label purity with in these subsets. This statistical significance of identifiers can de directly used for feature selection.

### 3.2 XGBoost Classifier Principle

In feature selection, to upgrade the effectiveness of producing new tree, XGBoost provides the best score for each and every element, shows significance of each element to prepare the model, and creates another tree with gradient direction [10, 11]. This factual hugeness of highlights can be straightforwardly utilized for feature selection.

XGBoost accomplishes exact classification by emphasis by ascertaining the frail classifier. It supports frail classifier and regression and is appropriate for building up regression model [12].

$$\widehat{XB} = \sum_{k=1}^k f_x(X_i) \quad \text{where } f_k \in E \tag{8}$$

K, represents the number of trees, E denote all possible regression trees, f denotes unambiguous regression tree. The objective function may have of two parts. First one is loss function (training error) and second one is normal constraint for tree.

$$f(\Theta)^t = \sum_i^n L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \omega(f_k) \tag{9}$$

At that point the target function for training can be composed as

$$f(\Theta)^t = \sum_i^n L(y_i, \hat{y}_i^{(t)} + f_i(X_i)) + \omega(f_k) \sum_{k=1}^t \omega(f_i) \tag{10}$$

### 3.3 Extra Trees Classifier Principle

Extremely Trees Classifier (aka Extra Randomized Trees Classifier) is a breed of ensemble learning technique that summarizes the outcome of different non-related decision trees cumulatively in a “forest” to targets its classification result. This is same as Random Forest Classifier and only differentiates in a way of generating the decision trees in the forest.

Every Decision Tree in the Extra Trees classifier forms by using actual dataset. At each trail node, every tree is allowed with a random specimen of n features from the actual data by which each decision tree must select the exact feature to split the data based on operational function (typically the Gini Index). This various specimen of features provokes to the creation of multiple non-related decision trees [13].

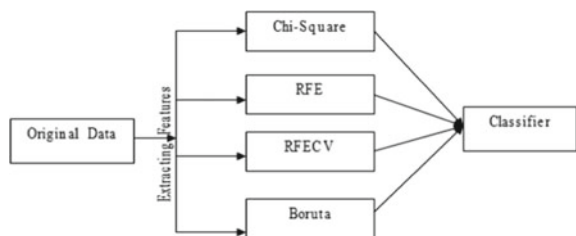
To achieve this extraction of features by using the forest structure, during the creation of forest, for each instance, the normalized mathematical criteria is used to take decision for the feature to split (Gini Index if the Gini Index is used in the construction of the forest) for computation. This procedure is referred as Gini Importance of the instances. To perform feature selection, each feature is ordered in lowered order according to the Gini Importance of each feature and the user selects the best n features from it [14].

## 4 Feature Selection Importance

In this proposed approach there is an ensemble method which combine the features extracted through various filters by feature extraction methods. For a specific type, if similar feature is preferred by all the selectors, then the instances is selected by using optimal approach [9].

In this work, the methods of feature preference used are Univariate Statistical Tests- Chi- Square, RFE and RFECV and Boruta see Fig. 1. As the data inside the dataset consists of Area Code, village code, population, Geographical area, no. of private and Government pre-primary, primary, Secondary and senior secondary schools in the area with its distance. To check whether the assets are sufficient particular area with in the specified range with the threshold value [15].

Fig. 1 Proposed framework for asset mapping





The features inside the dataset are independent and the feature selection is fully based on the target feature which is dependent. First 70% of data should be trained and remaining 30% of data should be tested. If the training dataset is imbalance then choose either majority class or minority class for selection of subset and apply different procedures of subset selection, and then combine different those subset to get optimal subset feature and then apply different classifiers for predicting accuracy [16].

Feature extraction will be mainly done on the basis of selecting relevant features and non-redundant features [17].

```

Algorithm for feature selection
-----
Input: D={X,L} // training set with n number of features.
           X={f1,f2,f3,...fn} and L labels.
X1           // Target feature(X1 ? X or X1={?})
⊖             // end the feature selection
Output: X1opt // an optimal subset
-----
Begin
Initialize:
Xopt = X1;
Ūopt = E(X1,A); // Evaluate the target based feature X1 by pruning A
do begin
Xg = generate(X); // subset generation
Ūopt = E(Xg,A); // subset calculation after a set of feature selection
If(Ū > Ūopt)
Ūopt = Ū
X1opt = Xg
repeat
End
return X1opt;
end,
    
```

**Lemma** Relevant features and non-redundant feature by removing bias.

**Proof** The feature selection method selects a feature f<sub>n</sub> and generates a subset (X<sub>g</sub>) and a subset computational measure A. This approach choses an optimal subset that good suits to learn an model.

### 5 Forecasting Model

Algorithm for Classification, attainment is calculated by the Confusion Matrix which contains data about the predicted and the actual class [18] (Table 1).

The true positive rate (TPR) is the values of features that is exactly diviner to be pragmatic, separated by the total no. of features in which the actual class is pragmatic

$$TPR = \frac{TPR}{TP + FN} \tag{11}$$

**Table 1** Prediction table for classification

Actual	Predicted	
	Positive class	Negative class
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

**Table 2** Feature Selection methods used in asset mapping

S. No.	FS method	Explain
1	CHI	Chi-square test
2	RFE	Recursive Feature Elimination
3	RFECV	RFE with cross validation
4	Boruta	Collecting best features

The false positive rate (FPR) is the total number of features that are in correctly diviner to be pragmatic, separated by the total no. of features in which the actual class is non pragmatic

$$FPR = \frac{FP}{FP + TN} \quad (12)$$

## 6 Experimental Design

This work elevates an ensemble feature preference based on XGBoost classifier. The dataset is taken from government of Andhra Pradesh, which consists of 32 features sample with 1494 dimensions (Table 2).

## 7 Experimental Result

To pick the features from the data set, the implementation is done in python programming by using different feature selection procedures on dataset with various classifiers.

Firstly, a dataset is chosen to select the features by using Chi-Square, RFE, RFECV and Boruta and then summarizing the features. Then the predictive model is constructed by XGBoost, Random Forest and Extra Trees Classifier. The attainment of the prediction model is evaluated by using Confusion Matrix. Finally, monotonous attempts are conducted on dataset to compare the performance measure. Finally by observing different bar graphs and accuracy of different classifiers for the dataset, XGboost provides high accuracy in extracting the features.

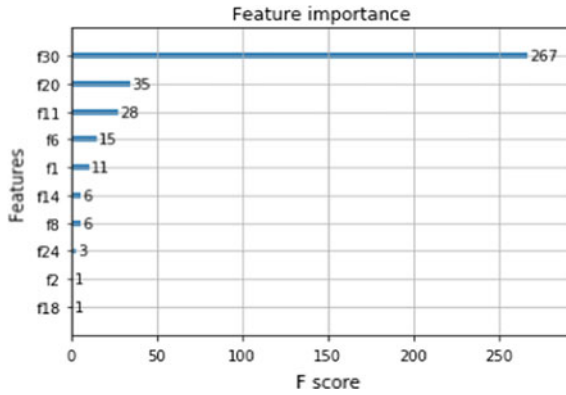


Fig. 2 Extracting the features from XGBoost classifier

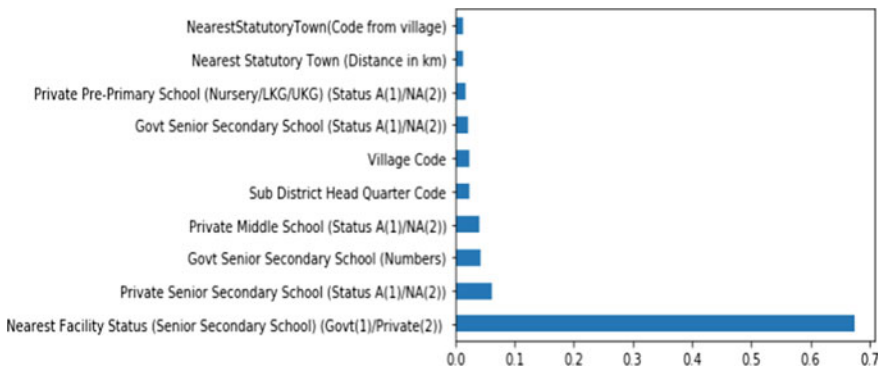


Fig. 3 Extracting the features by Chi-Square and prediction by Extra Tree classifier

### 7.1 Classification Results

See Figs. 2, 3, 4, 5, 6, and 7.

### 7.2 Comparative Analysis

Among these three classifiers, XGBoost gives best feature based on the target feature. Models were created based on the instances predict higher AUC values. The confusion matrix is for, the number of perfect predictions for each class and the number of imperfect predictions for each class, will be predicted based on the performance model.

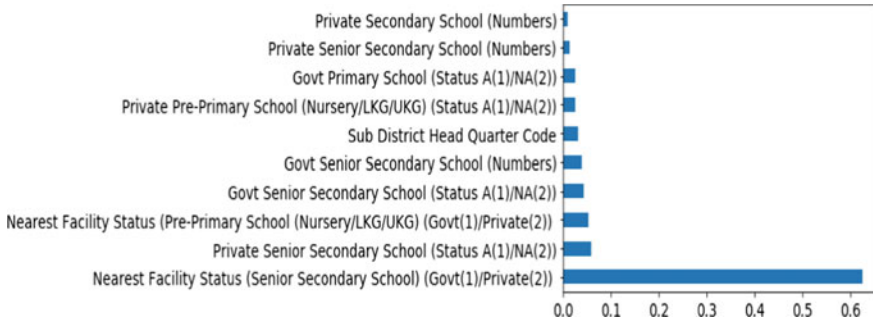


Fig. 4 Extracting the features by Chi-Square and prediction by RForest classifier

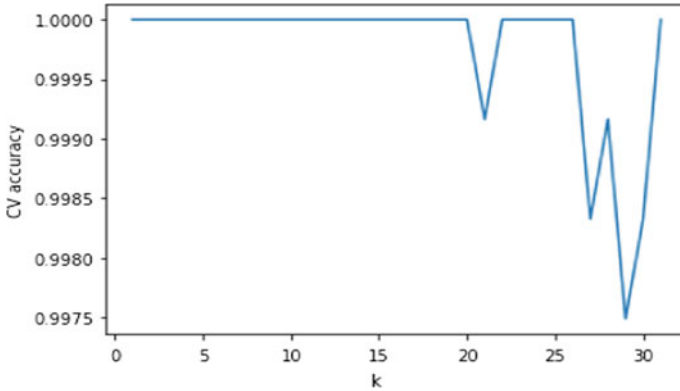


Fig. 5 Extracting the features by RFECV and prediction by Random Forest classifier

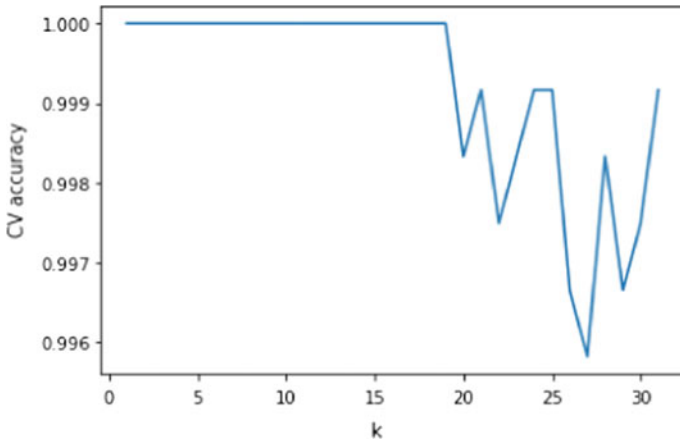


Fig. 6 Extracting the features by RFECV and prediction by Extra Tree Classifier

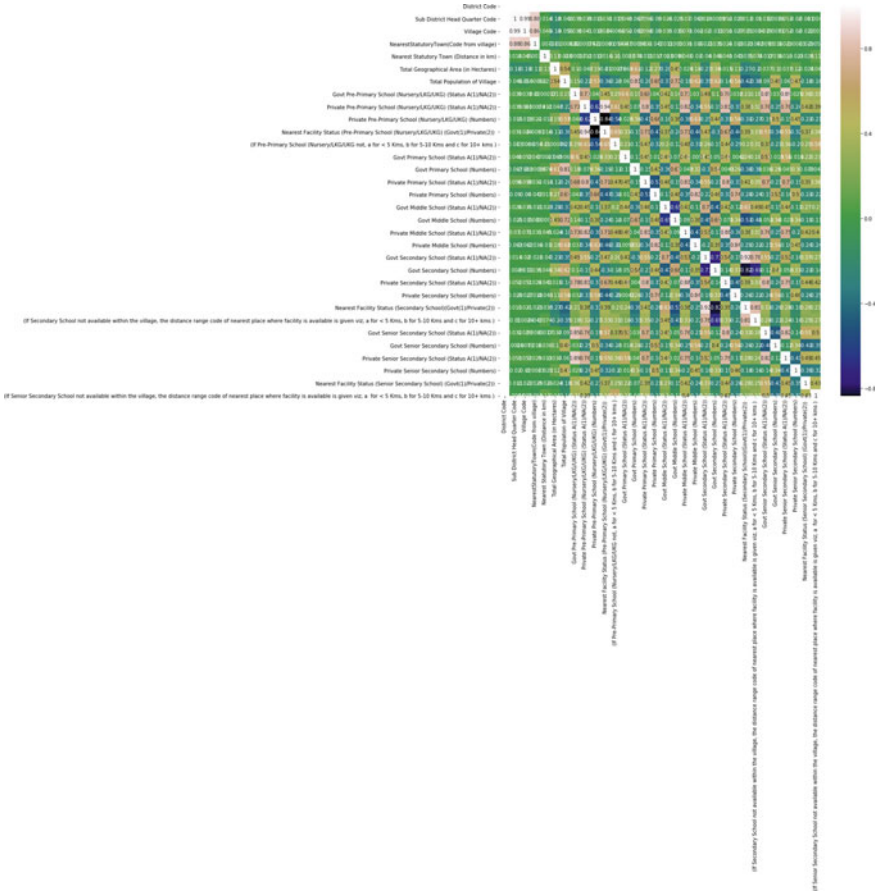


Fig. 7 Confusion matrix asset mapping dataset

### 8 Conclusion

This paper takes data set from government of Andhra Pradesh to test that the assets are sufficient in the area for the people residing in that community. To find the prediction accuracy-extract features and check the threshold intervals. For this process, data preprocessing is done and then followed by feature extraction and then classification models are constructed for prediction. Among them, XGBoost classifier provides high prediction accuracy.

## References

1. I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)
2. L. Jiang, S. Jiang, Q. Yu, Feature selection method based on sorting integration in software defect prediction. *J. Chin. Comput. Syst.* **39**(7), 36–40 (2018)
3. G. Chandra Shekar, F. Sahin, A Survey on Feature Selection Methods (Pergamon Press); C. Du, C. Zhou et al., Application of ensemble feature selection in gene expression data. *J. Shandong Univ. Sci. Technol. (Nat. Sci.)* **38**(1), 85–90 (2014)
4. A. Bidgoli, M.N. Parsa, A hybrid feature selection by resampling, chi squared and consistency evaluation techniques. *Eng. Technol.* **6**, 276–285 (2012)
5. J. Yang, Study on ensemble feature selection of biomics data (2017)
6. W. Altidor, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, Ensemble feature ranking methods for data intensive computing applications, in *Handbook of Data Intensive Computing* (Springer, Berlin, 2011), pp. 349–376
7. A.Y. Zomaya, Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. in *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Post Processing of Biological Data* (2017)
8. V. Bolón-Canedo, N. Sánchez-Marroño, A. Alonso-Betanzos, Data classification using an ensemble of filters. *Neurocomputing* **135**, 13–20 (2014)
9. S.D. Bay, Combining nearest neighbor classifiers through multiple feature subsets, in *ICML*, vol. 98 (Citeseer, 1998), pp. 37–45
10. N. Hoque, M. Singh, D.K. Bhattacharyya, EFS-MI: an ensemble feature selection method for classification
11. W. Hu, K.S. Choi, Y. Gu, S. Wang, Minimum–maximum local structure information for feature selection. *Pattern Recogn. Lett.* **34**(5), 527–535 (2013)
12. L. Torlay, M. Perrone-Bertolotti, E. Thomas, M. Baciú, Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform.* **4**, 159–169 (2017). <https://doi.org/10.1007/s40708-017-0065-7>
13. M. Ali, R. Ali, W.A. Khan, S.C. Han, J. Bang, T. Hur et al., A data-driven knowledge acquisition system: an end-to-end knowledge engineering process for generating production rules. *IEEE Access* **6**(99), 15587–15607 (2018). <https://doi.org/10.1109/ACCESS.2018.2817022>
14. M. Ali, UFS—Unified Features Scoring Code, version 1.0 (2017). Accessed 4 Apr 2018. Available online <https://github.com/ubiquitous-computing-lab/Mining-Minds/blob/master/knowledge-curationlayer/DDKAT/src/main/java/org/uclab/mm/kcl/ddkat/dataselector/FeatureEvaluator.java>
15. V. Bolón-Canedo, N. Sánchez-Marroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification. *Pattern Recogn.* **45**(1), 531–539 (2012)
16. A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning. *Artif. Intell.* **97**(1), 245–271 (1997)
17. S. Abdullah, N.R. Sabar, M.Z.A. Nazri, M. Ayob, An exponential monte-carlo algorithm for feature selection problems. *Comput. Ind. Eng.* **67**, 160–167 (2014)
18. O. Osanaiye, H. Cai, K.K.R. Choo, A. Dehghantanha, Z. Xu, M. Dlodlo, Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *EURASIP J. Wirel. Commun. Netw.* **2016**(1), 130 (2016). <https://doi.org/10.1186/s13638-016-0623-3>

# Land Site Image Classification Using Machine Learning Algorithms



G. Nagalakshmi, T. Sarath, and S. Jyothi

**Abstract** In contemporary improvements remote sensing and GIS technologies offer great tool for mapping and detecting amend in land use/land cover (LULC). LULC is a vital aspect in perceptive the dealings of individual deeds with the environment and thus it is required to replicate alteration. The primary objective was mapping process with classification scheme and procedural steps for interpretation so as to maintain standard operational procedures. As of now satellite hyperspectral image provides information but this is not sufficient to classify the living areas. So, to get the sufficient information about land use data LANDSAT images has been taken. Here Tirupati area is selected to classify, because fast growing city in Andhra Pradesh which is a pilgrim center suited at Rayalaseema region with in Chittoor district. Tirupati is divided in rural and urban area. Here to classify the image Machine Learning Algorithms like support vector machine, K-NN algorithms are applied and to know which algorithm is best for classification is compared with accuracy.

**Keywords** Remote sensing · LANDSAT images · Hyperspectral image · Support vector machine · K-NN algorithm

## 1 Introduction

The land cover/land use survey was first done by L.D. stamp in 1930 at Great Britain. In 1940 kelling applied land classification system to realize land use effort for development purpose by using aerial photographs [1]. However, the LULC information

---

G. Nagalakshmi (✉)

Department of CS, Rashtriya Sanskrit Vidyapeetha, Tirupati, India

e-mail: [agnl.lakshmi@gmail.com](mailto:agnl.lakshmi@gmail.com)

T. Sarath

Department of CSE, Siddhartha Institute of Science and Technology, Puttur, India

e-mail: [sarath.pallapalli1@gmail.com](mailto:sarath.pallapalli1@gmail.com)

S. Jyothi

Department of MCA, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, India

e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,

Learning and Analytics in Intelligent Systems 15,

[https://doi.org/10.1007/978-3-030-46939-9\\_19](https://doi.org/10.1007/978-3-030-46939-9_19)

and classification with remote sensing data was formulated in USA by 1971. The committee composed and implemented in the geographical analysis of the U.S. The land cover/land use classification [2] system which tin efficiently utilize detour and high altitude remote sensor data ought to rally the follow criteria

1. The accuracy of identify LU and LC category since satellite data may have 85%.
2. The correctness for numerous category must be about alike.
3. The results should be obtained from different sensing images of different time.
4. The classification can be appropriate more than wide area.
5. The different category must allow plants and other type of LC to be worn as alternates for bustle.
6. Aggregation of categories must be potential.
7. Various use of land must be recognized while feasible.

USGS formulated land cover/land use classification system for use amid remote sensing [3] data. The major classes are as follow: metropolitan or residential land, metropolitan or residential land farming land, variety land, jungle land, water, wetland, barren land.

In May 2006 national remote sensing center devised LU/LC classification method meant for utilize with satellite data in INDIA. The primary objective was mapping process with classification scheme and procedural steps for interpretation so as to maintain standard operational procedures. The classes are as follows: build up, farming land, jungle, wastelands, wetlands, water bodies etc.

Classification of LCLU is a criterion job in remote sensing in which every picture pixel is assign a class label signifying the object surface. A land use can enclose diverse land cover essentials to figure intricate structure. The information regarding the LULC is frequently store in geospatial databases, normally acquire and maintain by nationwide mapping agency. The main aim of LCLU is to describe the represented physical land type or how a land area is used. To full fill the above mentioned data firstly satellite images must be transformed into structured semantics. By seeing satellite images, observes different size and shape. Some may be readily identifiable while other may be not depending on our individual perceptions and experience. When identified our target seen on image the information is communicated with others, we are practicing interpretation.

A wide range of studies has been done on traditional parametric classifications [4] for remote sensing images [5] and to produce high accuracy for the composite data with a high dimensional [6] feature liberty machine learning algorithms are very much used.

## ***1.1 Machine Learning Algorithm***

Machine Learning (ML) is a rule, improperly substituted with Artificial Intelligence (AI) [7], but ML is a sub field/kind of AI. ML ordinarily named as prognostic analytics, or prognostic modelling. In 1959 Arthur Samuel a computer scientist in America



defined machine learning as a “computer’s capability to be told whereas not being freeway programmed” on its simplest. In ML programming algorithms are used for the analyse which receive input file and predict output values at intervals in appropriate vary. The novel information is nurtured to these procedures, they study and optimise their operation to progress concert, initial ‘intelligence’ over time. ML algorithms are differentiated in four forms there are supervised, semi-supervised, unsupervised and reinforcement.

### 1.1.1 Supervised Machine Learning Algorithm

A supervised machine learning algorithmic programme (as opposite to an unsupervised machine learning algorithm) is one that depends on targeted input data to be told to operate that produces associate degree applicable output when given new unlabeled data. Imagine a computer may be a kid, we have a tendency to its supervisor (e.g. blood relation, custodian, or educator), and we desire the kid (computer) to learn how a pig look like. We will show the kid many completely different footage, some of which are pigs and the rest could be pictures of anything (cats, dogs, etc.). When we see a pig, we shout “pig!” while it’s not a pig, we have a tendency to shout “no, not pig!” once doing this many times with the Kid, we show them a picture and ask “pig?” and they can properly (most of the time) say “pig!” or “no, not pig!” depending on what the picture is. That is supervised machine learning.

### 1.1.2 Semi Supervised Learning

The Semi supervised learning method is somewhat similar to supervised learning, however it is a substitute mutually uses labeled and unlabeled data. Labeled data is vital info to facilitate significant tag so to facilitate the process tin realize the data, whereas unlabeled data shortages that info. By use this arrangement, machine learning algorithms are discovering to make unlabeled data.

### 1.1.3 Unsupervised Machine Learning Algorithm

Unsupervised learning may be a machine learning technique, wherever it doesn’t you are having to be complied with supervise method. However, the model wishes to work on its own to urge data. here it will mostly deal with unlabeled info. An unsupervised learning algorithm permits to execute a lot of complicated process chores linked to supervised learning. Even though, unsupervised learning is extra unpredicted associated through different natural learning methods.

For illustration, taken the case of a babe and her domestic dog. The babe is aware of the dog and identifies it. After some days the relatives came along with the dog and there are tried to play with the baby. The baby not seen that dog earlier. However, she acknowledges several options (using 4 legs for walking, 2 ears and 2 eyes)

are as similar to her household dog. By that she recognizes as a dog. This type of identification is called unsupervised learning, however you are not trained but you have to discover it from info (here taken dog as a case.) in supervised learning, the relatives would tell the babe that it is a dog.

### **1.1.4 Reinforcement Learning**

Reinforcement learning emphasizes on restricted learning process, where a Machine Learning Algorithm (MLA) is providing a set of activities, parameter and finish value. By momentous policy, the MLA strive to examine diverse choice, budding, monitor and estimate every outcome to resolve which is an ideal. Reinforcement learning imparts the machine assessment plus fault. It studies from precedent experience and initiates to adjust the situation in retort the case to understand the simplest possible outcome.

## **2 Proposed Study**

In this paper, to analyses land cover change for Tirupati city we have taken shape file for classification [8]. Form that shape file latitude and longitude is taken to search for the LANDSAT image. For that firstly we study about Tirupati region and LANDSAT satellite image.

### **2.1 About Tirupati**

Tirupati is a prehistoric divine city placed within the south-eastern part of Andhra Pradesh. It's referred as the dwelling of God Venkateswara and is positioned within the Chittoor district of the state. Positioned at the foothills of the Eastern Ghats, reachable many major cities like Bangalore, Vijayawada, Hyderabad and Chennai concerning. Tirupati [9] is thought for Tirumala Venkateswara Temple that is one amongst the foremost necessary pilgrim sites in India and attracts several devotees every year. The temple is found at the Tirumala Hills that are one amongst the oldest rock mountains. It's believed that the temple even had devotees United Nations agency belonged to the good erstwhile dynasties just like the Pallavas, the Pandyas, the Cholas and therefore the kings of Vijayanagara. Geographically, town is found at 13° 37' N/79° 25' E. As per 2011 Census as shown in Fig. 1, the population of Tirupati is roughly 2.9 lakh. It experiences moderate winters and extreme summers. Telugu is its official language; however, Tamil is boot wide spoken at intervals town due to its proximity to Tamil Nadu.

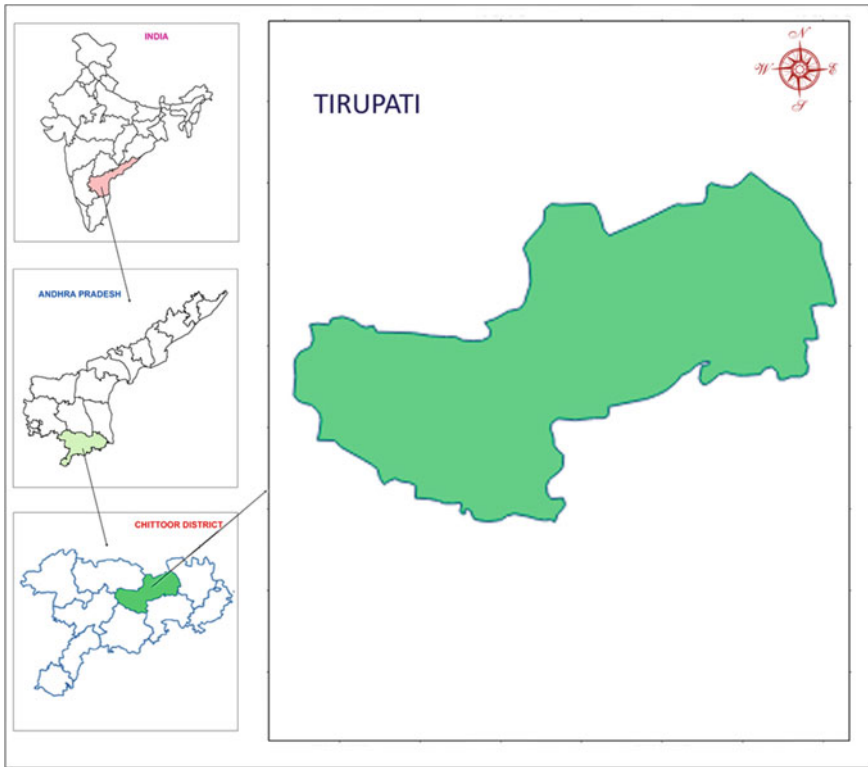
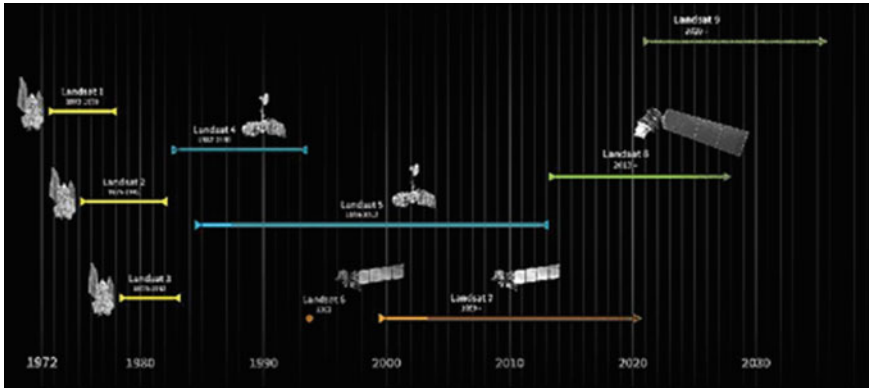


Fig. 1 Selection of Tirupati city from shape file

## 2.2 LANDSAT

The LANDSAT is a library of Globe imagery. Which is a series of satellites that become the treasure of nation. The LANDSAT has to commit and confirm the continuity of data. This can be referred as LANDSAT-8 in first it is titled as LANDSAT Data Continuity Mission (LDCM). The collected scene of LANDSAT might have a very tiny price. However, in past 100 years ago there is a burning question how to find the place where we are. This can be find by satellite data which is exists in globe for access the data. There are 8 LANDSAT satellites are present. Among them one satellite is not created in orbit but remaining seven satellites did. LANDSAT-9 may launched in 2023. It begins in 1972 with Landsat-1 as shown in Fig. 2. Formerly, it was titled as *Earth Resources Technology Satellite (ERTS)* and then it has been modified to *LANDSAT*. Currently, LANDSAT-8 is entitled as *LANDSAT Data Continuity Mission* however the public generally states it as just LANDSAT-8. Among the net observers to appear at close to period assortment of satellite imagery that currently viewed by the USGS Earth. This novel attribute displays *wherever* liberty LANDSAT is point by stream data.



**Fig. 2** Legacy of LANDSAT satellite. *Source* <https://directory.eoportal.org/web/eoportal/satellite-missions/l/landsat-9>

### 3 Methodology

#### 3.1 Support Vector Machine

A Support Vector Machine(SVM) builds a hyperplane or set of hyper planes in a high or endless dimensional space, used for classification. The hyperplane has the major expanse to nearby training data point of any class, commonly grader margin lowers the generalization fault of the classifier. The SVM uses binary classifier approach, which can handle more input data very resourcefully. Performance and accuracy depends upon the selection of hyperplane and parameter of kernel.

In constant quantity classification, the goal is to typify the distinctive aspect whole morals or dispersal of every category. In distinction, SVM emphases completely on the guidance sections that are next in the feature liberty to the best margin among the categories [10]. These samples are referred to as support vectors plus provides the tactic its title. The purpose in SVM [11] is seek out the best margin that exploits the parting, or edge, among the support vectors. The SVM classifier is integrally binary, characteristic one margin among two kinds. Though, this dispute is prevented by constantly relating to the classifier for each attainable mixture of categories, although this will entail that interval ought to rise exponentially because the variety of categories will increase SVMs [12] were formerly planned to spot a linear category boundary (i.e. a hyperplane). This drawback was self-addressed done by the prediction of the feature liberty to the next aspect, below the idea the linear boundary could occur during a higher dimensional feature liberty. This prediction to the next spatial property is thought because the kernel habit. Around several attainable kernels and every kernel could take a special set of needed user-specified limits. Communal kernels utilised in remote sensing are polynomial kernels and therefore the Radial Basis Function (RBF) kernel [13]. For categories that are inherently not severable, the choice boundaries are often thought to be consuming a soft-margin. This suggests

that drill category models are permitted on the incorrect facet of the edge, though a value, specific by the operator concluded the C parameter, is applied to those ideas. Thus, difficult C values can end in lot of complicated choice boundary, and fewer generalizations.

### 3.2 *K-NN Algorithm*

The k-Nearest Neighbour algorithm is not like other remaining classifier algorithm; therein the data is not trained to supply the model. Instead, every unknown sample is directly related beside the first training data [14]. The unidentified model is allotted to the foremost collective category of the k training models that are adjoining within the feature authority to the mysterious model. A squat k can thus turn out an awfully complicated call boundary; a better k can end in bigger generalization [15]. As a result of a trained model isn't created, k-NN classification would be estimated for bigger properties because the ranges of drill models will increase.

## 4 Experimental Analysis

To analyse the land cover/use classification India shape file is taken from that by selecting the Chittoor district, in that Tirupati city region is selected for our classification as in Fig. 1 this selection is done in python. From Tirupati region it is possible to get the latitude and longitude of the region, by using that geographical data LANDSAT images are possible to download which contains LANDSAT version, path, row, year, month, day, process, archive version of every image with band widths and this image is stored in.tiff file format. In LANDSAT 2, 3, 4, 5, 6 bands are chosen for classification and it form a grey image as shown in Fig. 3. The grey image is converted to RGB image from that classify the land cover.

Now, NDVI is applied to RGB LANDSAT image and the image is shown in Figs. 4 and 5 and then applied machine learning algorithms. Firstly, by applying the support vector machine algorithm are used to apply training data which called ROI (Reason of Interest) to find buildings, water bodies, non-build land etc., by this it is possible to classify land cover for our image, then k-means algorithm is applied for training the data for the classification of image. By comparing both the methods it is possible to evaluate the accuracy [16] of each method where SVM the accuracy of 0.9383 and k-means got the accuracy of 0.78. By that accuracy level it says that SVM classifier has the good accuracy than k-means. In LANDSAT image it will not select exact location so; hence it is derived particular latitude and longitude location of Tirupati location as shown in Fig. 6.

Fig. 3 LANDSAT image

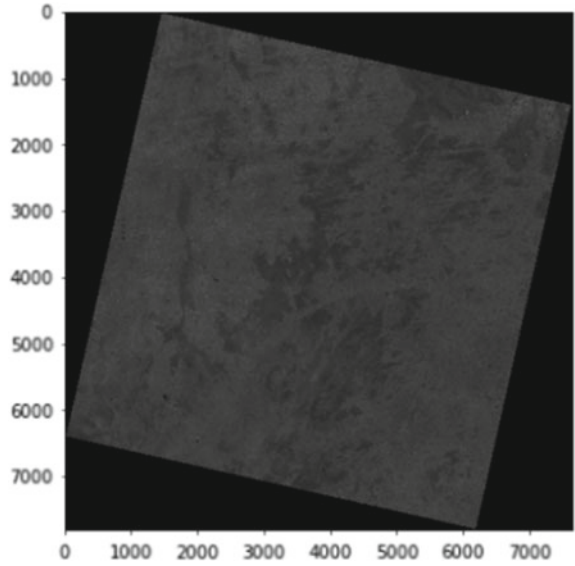
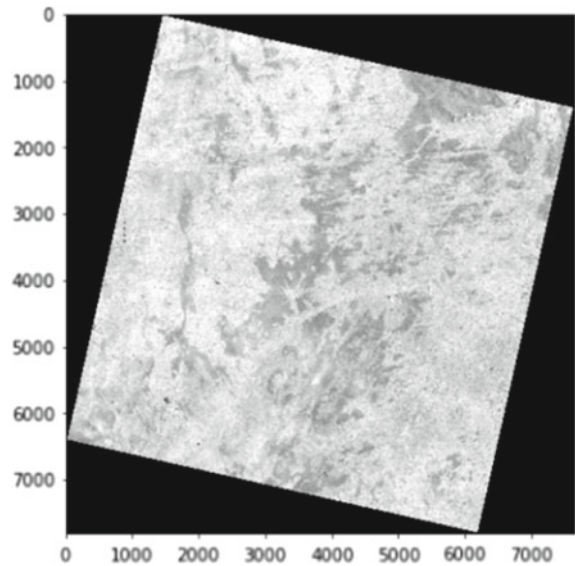


Fig. 4 RGB image



## 5 Conclusion

In this paper, study of Tirupati city region, and the shape file is used for our classification from India shape file. In that shape file latitude and longitude is taken to download the LANDSAT image for that particular region. To LANDSAT image

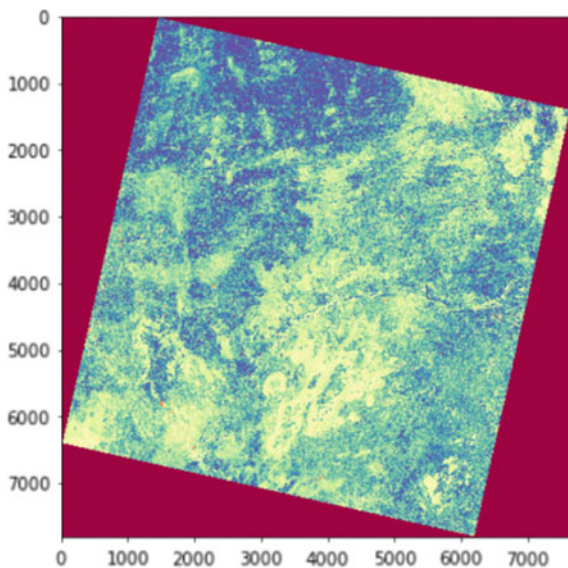


Fig. 5 NDVI image

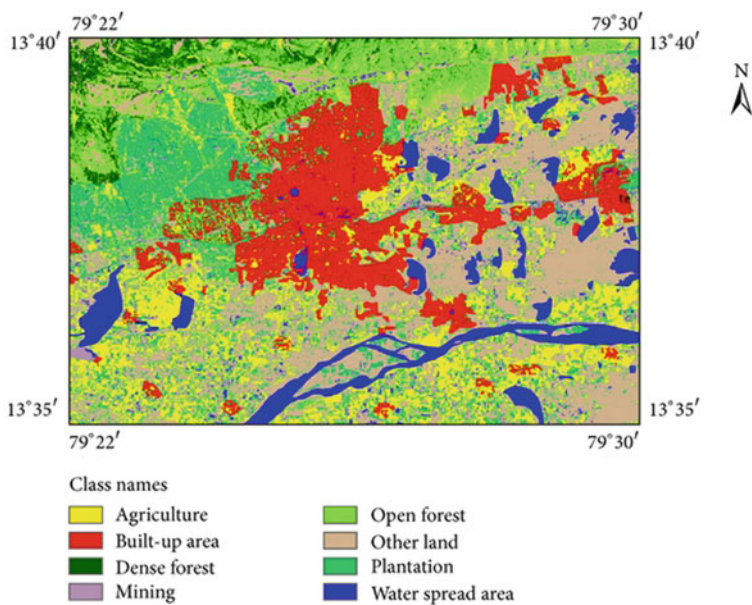


Fig. 6 Land cover/use classification of Tirupati

RGB, NDVI is applied and then machine learning algorithms like SVM, K-NN algorithms is realistic to classify the land cover of LANDSAT image, and then particular region of Tirupati area is selected to classify buildings, water bodies, non builds etc., for these algorithms accuracy was found to classify which algorithm is best to for image. By that it specified that SVM algorithm is more accurate than k-means algorithm. This can be found by applying any kind of machine learning algorithm and also deep learning algorithm.

## References

1. A.P. Gautam, E.L. Webb, G.P. Shivakoti, M.A. Zoebisch, Land use dynamics and landscape change pattern in a mountain watershed in Nepal. *Agric. Ecosyst. Environ.* **99**, 83–96 (2003)
2. J. Peng, Y.L. Wang, J.S. Wu, J. Yue, Y. Zhang, W.F. Li, Ecological effects associated with landuse change in China's southwest agriculture landscape. *Int. J. Sustain. Dev. World Ecol.* **13**, 315–325 (2006)
3. J.R. Anderson, E.E. Hardy, J.T. Roach, R.E. Witmer, *A Land Use and Land Cover Classification System for Use with Remote Sensor Data* (United States Government Printing Office, Washington D.C., 1976)
4. T. Subhash, S. Akhilesh, S. Seema, Comparison of Different image classification techniques for land use land cover classification: an application in Jabalpur District of Central India. *Int. J. Remote Sens. GIS* **1**(1), 26–31 (2012)
5. T. Sarath, G. Nagalakshmi, S. Jyothi, A study on hyperspectral remote sensing classification. *Int. J. Comput. Appl.* (2015). 0975-8887. *International Conference on Information and Communication Technologies* (2014)
6. T. Sarath, G. Nagalakshmi, An land cover fuzzy logic classification by maximum likelihood. *Int. J. Comput. Trends Technol. (IJCTT)* **13**(2) (2014). [arXiv:1407.4739](https://arxiv.org/abs/1407.4739)
7. K. Sree Divya, P. Bhargavi, S. Jyothi, Machine learning algorithms in big data analytics. *Int. J. Comput. Sci. Eng.* **6**(1), 63–70 (2018). E-ISSN. 2347-2693
8. A. Rajitha, P. Bhargavi, S. Jyothi, Hyperspectral image classification using soft computing techniques: a review. *Int. J. Comput. Appl.* **182**(11), 18–25 (2018). ISSN. 0975–8887
9. M. Sirish Kumar, B. Kavitha, S. Jyothi, G. Nagalakshmi, Land use/land cover of Tirupati area for agriculture land classification: a study. *Int. J. Eng. Res. Comput. Sci. Eng. (IJERCSE)* **5**(4) (2018)
10. V. Vapnik, C. Cortes, Support-vector networks. *Mach. Learn.* **20**, 273297 (1995). <https://doi.org/10.1007/bf00994018>
11. G. Nagalakshmi, T. Sarath, An SVM fuzzy logic classification for land cover. *Int. J. Manage. Technol. Eng.* ISSN NO: 2249-7455 (2018)
12. C. Huang, L.S. Davis, J.R.G. Townshend, An assessment of support vector machines for land cover classification. *Int. J. Remote Sens.* **23**(4), 725–749 (2002). <https://doi.org/10.1080/01431160110040323>
13. M. Pal, P.M. Mather, Support vector machines for classification in remote sensing. *Int. J. Remote Sens.* **26**(5), 1007–1011 (2005). <https://doi.org/10.1080/01431160512331314083>
14. N.S. Altman, An introduction to Kernel and Nearest-Neighbor nonparametric regression. *Am. Stat.* **46**(3), 175–185 (1992). <https://doi.org/10.1080/00031305.1992.10475879>
15. F. Maselli, G. Chirici, L. Bottai, P. Corona, M. Marchetti, Estimation of Mediterranean forest attributes by the application of K-NN procedures to multitemporal landsat ETM+ images. *Int. J. Remote Sens.* **26**(17), 3781–3796 (2005). <https://doi.org/10.1080/01431160500166433>
16. M. Pal, G.M. Foody, Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**(5), 1344–1355 (2012). <https://doi.org/10.1109/jstars.2012.2215310>



# Decades of Research and Advancements on Fabrication and Applications of Silk Fibroin Blended Hydrogels



Sufia Sultana, D. M. Mamatha, and Syed Rahamathulla

**Abstract** The Hydrogels are tunable three dimensional polymer network attractive for their rich hydrophilicity along with structural similarity with the extracellular matrix that provide a cell proliferation and facilitate rapid cell to cell communication. These Hydrogels are largely focussed by many researchers in the field of medicine due to its capacity to act as scaffold for tissue regeneration, as injectable Hydrogel for sustained drug delivery, in encapsulation of the enzymes and many more. These are prepared by either natural or artificial polymer or both and the nature of selection of polymer depends on its functional characteristics. In this review we reminisce different fabrication techniques and applications of Silk Fibroin (SF) blended with other polymers. SF acts as an attractive class due to excellent mechanical strength, biocompatibility that doesn't trigger any adverse immunological reaction and manageable biodegradation; in addition this Silk biomaterial has been used for suturing from past many centuries.

**Keywords** Biomaterial · Protein · Silk Fibroin (SF) · Silk blended Hydrogel · Biomedical

## 1 Introduction

Hydrogels are three dimensional polymeric networks that are hydrophilic and thus swells and abstains water in reverse direction. The nature of highly responsive to environmental stimuli like temperature change, variation in the pH and ionic strength made them excellent material for biomedical applications [1], especially in the application that are water based or soft material [2]. Wichterle and Lim first time created the cross linked 2-hydroxyethyl methacrylate (HEMA) Hydrogels that were used as drug carrier, as neoplasm and in osteoporosis [3]. Structurally Hydrogel are classified

---

S. Sultana (✉) · D. M. Mamatha · S. Rahamathulla  
Sri Padmavati Mahila Viswavidyalayam (Women's University), Tirupati, India  
e-mail: [sufiabitech@gmail.com](mailto:sufiabitech@gmail.com)

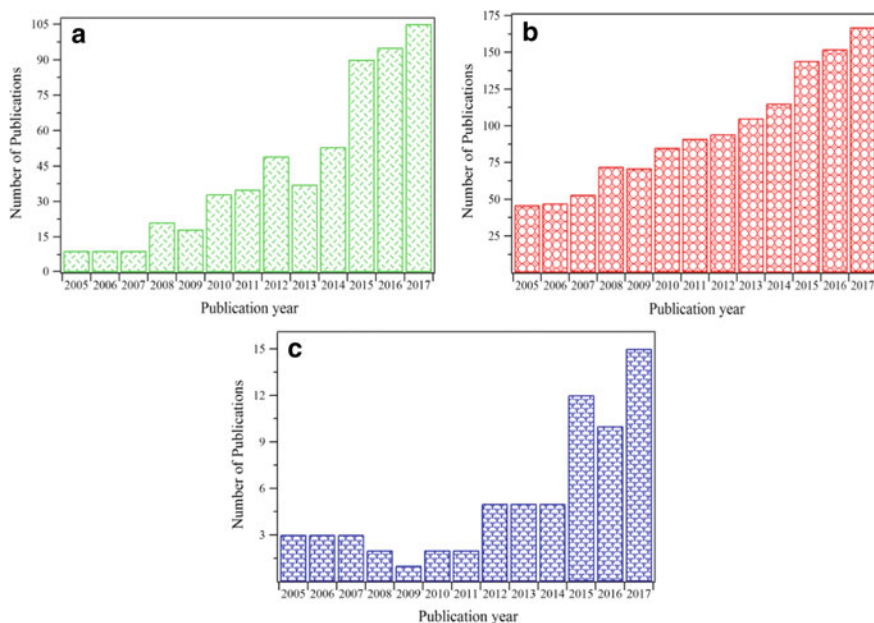
D. M. Mamatha  
e-mail: [prof.mamatha@gmail.com](mailto:prof.mamatha@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_20](https://doi.org/10.1007/978-3-030-46939-9_20)

into amorphous, oxygen bounded molecules, semi crystalline and super molecular structure or hydrocolloidal aggregations. They are prepared by either synthetic or natural polymers or in blended form carrying their own advantages and disadvantages. The synthetic polymers are attracted for their strength, manageable degradation kinetics with required molecular weight moieties [4]. Lack of cell recognition sites and probability to release toxic degradation products carry the adverse side [5]. On the other hand, natural polymers possess properties like bio renewability, biocompatibility, biodegradability and cell recognition. One of such natural protein polymer is Silk protein that date back many centuries ago as medical sutures. Later after enormous research in this modern science era found that Silk protein possess high mechanical strength, structural flexibility, biocompatibility, manageable biodegradability and provide the facility to immobilize the growth factors by modifying the amino acid side chain [6]. Moreover, Silk is FDA approved implant material for soft tissue repair and Serica is marketing the product [7]. Apart from that the Silk based sutures like Surusil<sup>®</sup>, Covidien, Suru and SofSilk<sup>™</sup>, and Silk based scaffolds like Seri<sup>®</sup> Surgical Scaffold, Allergan are at present in market. Research labs like AMSilk working on spider Silk, Immuno-Biological Laboratories on Transgenic Silkworm in producing recombinant human proteins, Banner Pharmacaps and Ekteino Laboratories researching on the Silk protein as drug delivery platforms and Vaxess on vaccine stabilization [8]. Moreover, the Silk proteins as biomaterial are in great demand as implants, scaffolds, drug delivery vehicles and medical photonics [9]. In order to be considered as a biomaterial, it should be sterilizable, biofunctional and biocompatible [10]. Where SF is versatile to sterilization with no changes under 120 °C [11]. The aim of this review is to illustrate the several preparation techniques used to develop Silk Hydrogel by blending with natural polymer, along with its applications. The increase in the number of publications on the Hydrogels and its blended forms indicate the demand in the current scenario explained in Fig. 1.

## 1.1 Structural Characteristic

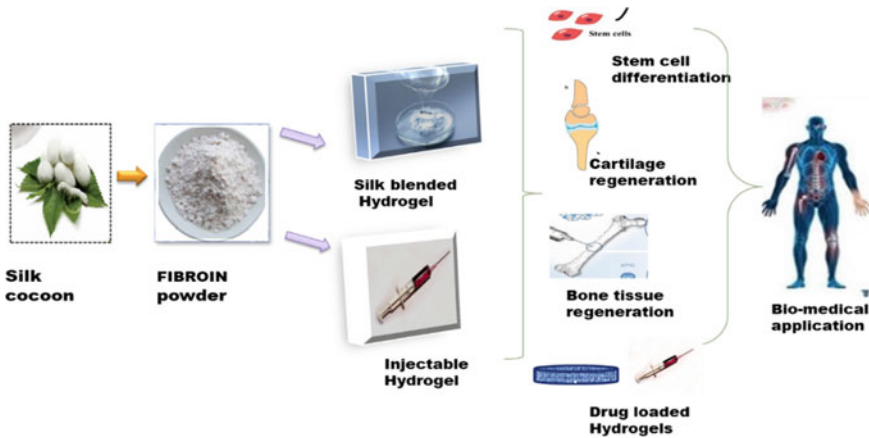
Silk is a naturally derived fibrous protein secreted in specialized epithelial cells of organisms such as spiders and Silkworm (*Bombyx mori*) where it protects the cocoon from environmental stress [12]. The spider protein though strong comparatively, domestication and economical profitability made Silkworm a preferred choice [13]. The Silkworm produces dual protein with outer Sericin and core Fibroin. The outer Sericin is removed by degumming due to its immunological effect. Silk Fibroin (SF) has heavy and light chain polypeptides of ~350 kD and ~25 K Da, respectively connected by a disulfide link and possess highly repetitive primary sequence. The SF also exists in amorphous and crystalline form. The crystalline component is of two conformations unstable-water soluble Silk I and stable-water insoluble Silk II. The repetitive hexa peptide Gly-Ala-Gly-Ala-Gly-X on heavy chain of SF gives it a crystalline conformation [14].



**Fig. 1** The number of peer-reviewed articles on the silk-blended hydrogels increases tremendously over the years. The utilization of silk hydrogels (topic, **a**), blend hydrogels (topic, **b**) and refine silk search results within blend hydrogels (topic, **c**) have been realized by searching the Web of Science database from 2005 to 2017 (this figure). This figure clearly shows exponential growth in the number of published articles over the last few years. This tremendous increase in publications is a general trend for any scientific topic publications. Although this statistical figure is not evidence, it is just an indication the growing interest on the silk-blended hydrogels. *Source* Web of Science database from 2005 to 2017

## 2 Preparation of SF Hydro Gels

The process involves silkworm cocoons, that were cut into small pieces followed by boiling in 0.02 M Sodium carbonate ( $\text{Na}_2\text{CO}_3$ ) solution in order to remove Sericin, glue like proteins and this degumming process is time-dependent. The obtained purified SF is dissolved in various chaotropic high concentration salts for e.g.  $\geq 9$  M LiBr,  $\geq 50$  °C or  $\geq 10$  M LiSCN,  $\geq 25$  °C or ternary systems containing alcohols for e.g.  $\text{CaCl}_2$ –water–ethanol,  $\geq 70$  °C. Subsequent desalting of SF solutions is carried out by performing the dialysis in 6000–8000 MWCO regenerated cellulose dialysis tubing [7]. The solutions were now autoclaved and centrifuged twice at 10,000 rpm for 25 min to remove formed SF aggregates and stored (maximum one week) at 4 °C. In the sonication-induced gelation method, 4% (w/v) SF solution was sonicated with Microson XL 2000 for 15 s at 20 W and in vortex-induced gelation, SF solution was mixed for 10 min at 3000 rpm using a vortex mixer (Fisher Scientific, Hampton, NH). All the solutions were incubated at 37 °C for a 12 h period [15].



**Fig. 2** Listed are few Bio-medical applications of Silk based hydrogel that are used in regeneration of tissue like Cartilages and Bones, in differentiation of Stem cells and many more

The Hydrogels are hydrophilic three dimensional polymer network resembles the extra cellular matrix. SF solution readily self assembles to form Hydrogel but its rate depends on protein concentration, pH, metal ions and temperature [16]. SF blended Hydrogel are formed with collagen, chitin, alginic acid, gelatin and hyaluronic acid. Each polymer carries with them both advantageous and disadvantageous characteristics, where the blending of these natural polymers can surpasses adverse properties thereby producing the excellent biomaterial for bio-medical application. Hydrogel represent a platform with diversity of biomedical applications like drug delivery systems, scaffolds for cell culture and bio-responsive systems that is represented in Fig. 2.

#### a. SF—Collagen Hydrogels

Collagen is fibrous protein that acts as scaffold for adhesion of cells and growth factors in the tissue. The pure collagen Hydrogel are readily prone to enzyme degradation [17]. A recombinant human-like collagen (RHLC)-SF scaffold for tissue engineering upon which fibroblasts are cultured resulted in the enhanced attachment and proliferation of fibroblasts. This hybrid scaffold possessed porosity above 90% suitable for tissue engineering [18]. In the similar study a rigid SF-collagen Hydrogel with nontoxic 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) as cross linker resulted in thermally and mechanically stiff substrate upon which vascular smooth muscle cells (VSMCs) cultured that showed a favourable cytocompatibility [19]. Chomchalao et al. [20] developed complex SF-based 3-D scaffolds blending with gelatine and collagen through freeze drying method. The Fibroin blended scaffolds possessed enhanced mechanical properties, promoted chondrocyte cell adhesion and proliferation due to presence of RGD a cell-recognition signal in the gelatine and collagen.

## 2.1 *SF-Chitin Derived Hydrogel*

The Chitin and Chitosan are cationic polysaccharide-biopolymers composed of randomly arranged N-Acetyl-D glucosamine and N-glucosamine units. Chitosan, derived from chitin through de-acetylation. The chitosan is rigid crystalline compound with three reactive sites suitable for modification and shows structural resemblance with the glycosaminoglycan of extracellular matrix thus used as scaffold for tissue regeneration. The non-economical nature, low solubility at physiological pH, difficulty in handling, over swelling ability and low grade mechanical characteristics limited its extensive use. Chitosan is used either as coated textile form or combined in PVA. The chitosan derivatives like N-O-carboxymethyl chitosan (CM-chitosan), quaternary chitosan (Q-chitosan), and carboxyethyl chitosan (CEchitosan) are found better than pure chitosan due to excellent anti-microbial activity in all pH conditions. On the other hand pure chitosan showed anti-microbial activity in only acidic conditions and lack of homogeneity with PVA. SF-chitosan Hydrogel using ultrasonication method developed where MC3T3-E1 cells were encapsulated and colonised. The study says this ultrasonication method is better than the chemical methods where in certain situations drugs or bioactive compounds encapsulated may denature due to the organic chemicals [21].

## 2.2 *SF-Alginic Acid Hydrogel*

Alginic acid is natural polysaccharide that consists of L-guluronic acid and b-D-mannuronic acid. Thomas for the first time used alginic acid sponges and gels along with its salt in wound dressing that resulted in enhanced wound healing [22]. In one of the study a SF-alginate Hydrogel scaffold developed to culture embryonic stem cells. The alginate for its quick gelation in presence of divalent salts and Silk Fibroin for its strength contributed in formation of physically stable Hydrogel. Non adherence property of Silk alginate Hydrogel was combated with the addition of laminin a basal membrane protein, promoted growth of embryonic stem cell. Ultimately the study says lack of stiffness (anisotropic nature) and in vivo gelation limited its further use [23]. Going in the same routes of biomimetry a nature inspired approach where use of organic polymer matrice for a controlled growth of inorganic minerals, Ming et al. [24] synthesized hydroxyapatite nano sized rod crystal on SF-sodium alginate to study the effects of factors like duration, temperature and pH. Hydroxyapatite is important mineral present in the teeth and bones. Controlled growth of hydroxyapatite on SF-sodium alginate serves as a path to study bone mineralisation. SF-sodium alginate nanofiber Hydrogel were also acted as scaffold for the generation of hydroxyapatite (HAp) crystal growth [25]. Biopolymer based hybrid/composites scaffold is more advantageous compared due to the enhanced tensile strength for tissue regeneration. Furthermore, in cell extracellular matrix are arranged in composite systems with its constitute functioning in a synergy.

### 2.3 *SF-Gelatin Hydrogel*

Gelatin is the product obtained by hydrolysis of the triple-helix structure of collagen into single-stranded molecule extracted from the bones, connective tissues and from skins of bovines and pigs. The gelatine has thermally reversible nature, at low temperature the random coil changes to triple helix whereas at elevated or body temperature reverse back to random coil thus results in dissolution of Hydrogel. SF doesn't possess the thermal-stimulus nature thus maintain stable structure even at the elevated temperature. So Gil et al. [26] prepared gelatin-SF Hydrogel to study morphological changes at both ambient (20 °C) and elevated temperature (37 °C). The induction of beta-crystal in the SF maintained the stable conformation in the gelatine-SF Hydrogel. Similarly, Hu and Kaplan D. L. synthesized mechanically strong and tunable methacrylated gelatine (GelMA)-SF photo-cross-linkable, tunable interpenetrating polymer network (IPN) Hydrogel [27]. More over the GelMA in the Hydrogel are more compatible with the micro fabrication techniques used in the regenerative medicine, promotes cell proliferation and migration in 2D [28] as well as 3D [29] structures.

### 2.4 *SF-Hyaluronic Acid Hydrogel*

Hyaluronic acid (HA) or hyaluronan is natural polymer consists of repeated units of 1, 4-linked D-glucuronic acid and 1, 3 N-Acetyl-D-glucosamine disaccharide. HA is anionic, non-sulphated, higher molecular weight glycosaminoglycan distributed widely in the extracellular matrix of connective tissues, epithelial tissues and neural tissues. Due to their major role in cell mobility and cell differentiation used in making scaffolds for biomedical application. Many scientists developed SF-HA Hydrogel using physical and enzymatic methods. The Silk contributes its mechanical strength and biocompatible nature while HA aids in enhanced water retention and stiffness. Hu et al. [30] developed HA encapsulated SF-HA Hydrogel by using sonication method. As HA is non cross linkable by physical method it is entrapped by beta-sheet crystallinity of SF where beta-sheet crystallinity of SF should be 26–34% and HA should be below 40%. In the same lines Elia et al. [31] prepared composite Hydrogel by entrapping electrospun SF in the poly (ethylene glycol)-diacrylate (PEGDA) cross-linked thiol group modified HA matrix for controlled release of therapeutic drugs. Entrapped reinforced SF provides strength and aids to target localized drug delivery. In a patented work Pavlovic et al. [32] developed SF-HA Hydrogel using 1-ethyl-3 [3-dimethyl aminopropyl] carbodiimide hydrochloride(EDC) in conjugation with hexa methylene diamine (HMDA). EDC activates and enables the carboxylic acid groups of HA that reacts with the cross linkers, resulting in formation of robust and sterilizable Hydrogel. They are used in dermal fillers, grafting procedure and also in transferring adipose tissue. Similarly, the lyophilized SF Hydrogel matrix (lyogel) with monoclonal antibodies embedded was developed that facilitated a sustain

released of the monoclonal antibody for over 38 days compared with 10 days release by Hydrogel. Similarly, Zhang et al. [33] fabricated PVA-HA-Silk composite/hybrid Hydrogel for orthogonal experiment to quantify the influence degree of three polymers PVA, Silk and HA on mechanical properties, amount of water, stress relaxation behaviour, elastic modulus and creep characteristic. PVA shown the strongest impact on the stress relaxation rate followed by Silk and HA. Whereas Silk has the strongest impact on the elastic modulus followed by PVA and HA. The experiment finally documented that PVA-HA-Silk composite Hydrogel with mass percentage of PVA 15%, HA 2.0% and Silk 1.0% is the best polymer ratio.

## 2.5 SF-Cellulose Hydrogel

Cellulose belongs to carbohydrate group made up of repeated anhydrous glucose moiety linked by a  $\beta$ -1,4 glycosidic bond found in plant and bacterial cell-walls, certain sea creatures such as tunicates and algae. At nanoscale, cellulose structure is advantageous over conventional cellulose fibres due to high surface area, aspect ratio, and Young's modulus [34]. A group of researchers [35] prepared SF-Cellulose Hydrogels using the concentrated lithium bromide (LiBr) aqueous solution as solvent. This method of preparation using the LiBr avoided the use of harsh chemicals used in conventional cellulose extraction. List below in tabular format are few pros and cons of Polymers (Table 1).

**Table 1** Advantages and disadvantages of polymers

Polymer	Advantages and disadvantages
Silk fibroin	Excellent polymer with good mechanical strength, biodegradability, biocompatibility. But brittle in dry state
Chitosan	A rigid crystalline compound with three reactive sites suitable for modification. But non-economical nature, low solubility at physiological PH, difficulty in handling, over swelling ability and low grade mechanical characteristics limited its extensive use
Alginic acid	A poly anionic copolymers replaces the role of beta-chitin in that generally occurs in many calcium based minerals to control crystallization to a specific space. But lack of stiffness and gelation limited its further use due absence of sufficient divalent ions/salts [22]
Collagen	The presence of RGD a cell recognition signal makes it suitable polymeric material [19]. However pure collagen are in general more rapidly degraded by enzyme and easily permeable to microbes compared to synthetic hydrogel
Hyaluroni acid	Even though hyaluronic acid possess enhanced stiffness and water retention capacity, it is rapidly degraded in the body due to the presence of hyaluronidase enzyme

### **3 Application of Silk Hydrogel**

In general Hydrogel have achieved tremendous attention owing to their applications in various fields like drug delivery, tissue engineering, selective transport, three-dimensional printing and biocatalysts.

#### ***3.1 Bone Tissue Regeneration***

Bone is highly vascularised, mineralized and tissue imparting; the unique mechanical properties to the bone tissue. Bone tissue is capable to self heal but in large tissue defect surgical intervention is compulsory. The repair methods like autografting and allografting are limited due to risks of donor-site morbidity which is resolved by bone scaffolds or bone filler [36]. Injectable Silk Hydrogel bone fillers showed significantly higher trabecular bone volume and rapid bone formation rate [37]. Similarly the injectable Silk Hydrogels loaded with both Vascular Endothelial Growth Factor (VEGF165) and Bone Morphogenic Protein 2 (BMP-2) growth factors resulted in large new bone formation than that loaded with either one or none [38].

#### ***3.2 Cartilage Regeneration***

Cartilage tissue composed of chondrocytes enclosed in extra cellular matrix. Its repair or regeneration by tissue engineering is of profound interest due to the incapacity of self-healing of damaged cartilage due to lack of vascularization, lymphatic networks, innervations and progenitor cells [39]. Chitosan/glycerophosphate (CS/GP) added with both chopped and electro spun Silk fibres acted as strong scaffold for cartilage regeneration [40]. The Silk Hydrogels compare well with that of agarose, which is a gold standard material for cartilage tissue engineering [41].

#### ***3.3 Stem Cell Differentiation***

Pluripotent stem cell has the potential to differentiate into any of the three germ layers. Stem cell therapy showed a way in the treatment of vascular diseases with the capability to regenerate vascular tissues both in vitro and in vivo. According to the 2009 USA report on Cardio vascular disease (CVD), CVD is a leading cause of death with 1:4 death ratios [42]. The current therapies such as autologous vascular bypass grafts are limited either by surgical intervention or the disease condition of the target explanted tissue. On Transforming growth factor (TGF-1)—SF Hydrogel, human mesenchymal stem cells (hMSC) differentiated into mature smooth muscle



within modest culture periods (72 h) [43]. TGF family is a potent regulator of cell proliferation and distribution and is also strongly related with vascular smooth muscle cell (vSMC) stem cells differentiation. Pluripotent stem cells cultured on gelatin-SF Hydrogel cross linked via genipin differentiated into epithelial ectodermal cells instead of neural ectodermal cells thus behaviour of stem cell depends on the substrate [44].

### ***3.4 Biocatalyst***

Hydrogels having a three-dimensional structure accommodate the enzymes thus acts as biocatalyst. The harsh preparation conditions and poor enzyme stability and activity rendered the use of synthetic Hydrogel and natural Hydrogels though possess good enzyme preserving activity, the weak mechanical properties, poor stability and elasticity limited their application in supporting enzymes. Here comes the significance of the Silk Hydrogel, the natural polymer with enhanced mechanical properties that effectively preserves the enzyme activity. The SF Hydrogel with carbonic anhydrase enzyme (CA) encapsulated acts as a promising biocatalyst for environmental friendly CO<sub>2</sub> sequestration with an excellent compressive modulus, high resiliency, elasticity and structural stability [45] but slow preparation time and encapsulation of undesired enzymes are listed drawbacks.

### ***3.5 Injectable Hydrogel***

The Injectable Hydrogels are solid before administration under proper shear stress they can turned into shear-thin and can subsequently flow, later which it retains back to solid state at the site of injection. The Hydrogel with poor implant fit, possible surgical wounds and susceptibility to infections possible leakage of liquid precursors to surrounding tissues. So recent trend is designing injectable Hydrogels that possess thixotropic properties [46]. Apart from this most of the drugs today are weakly soluble in water, so the need of biomaterial that enhance drug bioavailability. The high molecular weight injectable SF Hydrogel developed that possessed lower viscosity making it easy as injectable source, very slow-drug release and a zero-order rate when compared with polymers i.e. Pluronic F-127(PF-127) and chitosan and addition of SPH as adjuvate in PF-127 and chitosan Hydrogel caused further reduction of drug buprenorphine [47]. Further 95% of cell viability after injection observed in SF-hydroxypropyl cellulose (HPC) injectable Hydrogel with thixotropic properties [48]. The injectable Silk-hydroxyapatite (SF-HA) Hydrogel with thixotropic properties showed enhanced osteogenesis in regeneration of irregular bone defects [49].

### **3.6 Wound Dressing**

The Hydrogels mainly facilitate moist environment to the wound area which results in the reduction in the dryness and itching. SF-calcium alginate—carboxymethyl cellulose (CMC) Hydrogel though possess less adhesive capacity than the commercially available medical gauze (C) and Purilon Gel<sup>®</sup> (PG) but relatively has good cell growth and viability [50].

## **4 Conclusion**

At present, the field of biomedical sciences is focusing mainly on the use of biopolymer-based biomaterials implant systems than the synthetic implants like inert metal and plastic-based materials due to the necessity of replacing them from the patient body. This feature can be overcome by the use of degradable biopolymers and their corresponding composite. One among them is Silk based Hydrogel which are in great demand in the view of the Silk protein mechanical strength that imparts stiffness to the scaffolds, biocompatibility and biodegradation of SF overall aids in the suitability of Hydrogel in the various application of biomedicine. These Hydrogels will be promising lead for future biomaterial platforms. This review has provided key insights into the advanced Silk blended Hydrogel preparation technique carried by various experiments all over the world along with growing applications in biomedical engineering. These are proposed as ideal smart biological platforms due to spatiotemporal delivery of biochemical factors over the biochemical or physical stimuli. Even though SF biocompatibility, mechanical strength and blending of SF with other polymers along with controlled finely tuned processing techniques will certainly aid in the driving Silk Hydrogel to the future innovation only when certain limitation are paid attentions. They are enough though the natural Silk fibres are strong and flexible, the brittleness of Silk based material in dry state stressed on the blending of the natural polymer with the other polymers. On the other hand biochemical factors like pH, temperature and gelation time of SF during the Silk processing methods are to be considered as they decreases the biocompatibility of Silk. To gain full advantage out of SF blended Hydrogel it is compulsory to design and construct the Silk Hydrogel focussing mainly on the processing technologies.

## **5 Future Perspective**

In the fabrication of hybrid Silk biopolymer, much attention should be taken on the selection of appropriate polymers and fabrication techniques that yields the products with its most native properties, as regeneration of the polymers in making the various form results in loss of its original properties. Furthermore, on the final architectures

that will maintain the synergy in between constituents of the hybrid biomaterial. These features can only be gained by the interdisciplinary and collaborative efforts.

## References

1. A. Sood, M.S. Granick, N.L. Tomaselli, Wound dressings and comparative effectiveness data. *Adv. Wound Care (New Rochelle)* **3**(8), 511–529 (2014)
2. C.M. Kirschner, K.S. Anseth, Hydrogels in healthcare: from static to dynamic material microenvironments. *Acta Mater.* **61**(3), 931–944 (2013)
3. I. Gibbs, H. Janik, Review: synthetic polymer Hydrogels for biomedical applications. *Chem. Chem. Tech.* **4**, 297–304 (2010)
4. P. Gunatillake, R. Adhikari, Biodegradable synthetic polymers for tissue engineering. *Eur. Cells Mater.* **5**, 1–16 (2003)
5. S. Seo, C. Mahapatra, R. Singh, J. Knowles, H. Kim, Strategies for osteochondral repair: focus on scaffolds. *J. Tissue Eng.* **5**, 2041731414541850 (2014)
6. G.H. Altman, F. Diaz, C. Jakuba, T. Calabro, R.L. Horan, J. Chen, H. Lu, J. Richmond, L. Kaplan, Silk-based biomaterials. *Biomaterials* **24**(3), 401–416 (2003)
7. Serica Technologies. <http://www.sericainc.com/en-us/news/2009>
8. T. Yucel, M.L. Lovett, L. Kaplan, Silk-based biomaterials for sustained drug delivery. *J. Control Release* **190**, 381–397 (2014)
9. B. Kundu, N. Kurland, S. Banoa, C. Patrac, F. Engel, K. Vamsi, V. Yadavalli, S. Kundu, Silk proteins for biomedical applications: bioengineering perspectives. *Prog. Polym. Sci.* **39**, 251–267 (2014)
10. S. Bauer, P. Schmuki, K. Von Der Mark, J. Park, Engineering biocompatible implant surfaces part I: materials and surfaces. *Prog. Mater. Sci.* **58**, 261–326 (2013)
11. T. Furuzono, A. Kishida, J. Tanaka, Nano-scaled hydroxyapatite/polymer composite I Coating of sintered hydroxyapatite particles on poly (gamma-methacryloxypropyl trimethoxysilane) grafted Silk Fibroin fibers through chemical bonding. *J. Mater. Sci. Mater. Med.* **15**(1), 19–23 (2004)
12. Y. Zhang, P. Zhao, Z. Dong, D. Wang, P. Guo, X. Guo, Q. Song, W. Zhang, Q. Xia, Comparative proteome analysis of multi-layer cocoon of the Silkworm, *Bombyx mori*. *PLoS ONE* **10**(4), e0123403 (2015)
13. H.J. Jin, D.L. Kaplan, Mechanism of silk processing in insects and spiders. *Nature* **424**(6952), 1057–1061 (2003)
14. H. Yamada, Y. Igarashi, Y. Takasu, H. Saito, K. Tsubouchi, Identification of Fibroin-derived peptides enhancing the proliferation of cultured human skin fibroblasts. *Biomaterials* **25**(3), 467–472 (2003)
15. A. Zuluaga-Vélez, D.F. Cómbita-Merchán, R. Buitrago-Sierra, J.F. Santa, E. Aguilar-Fernández, J.C. Sepúlveda-Arias, Silk Fibroin hydrogels from the Colombian silkworm *Bombyx mori* L: evaluation of physicochemical properties. *PLoS ONE* **14**(3), e0213303 (2019)
16. U.J. Kim, J. Park, H.J. Kim, M. Wada, D.L. Kaplan, Three-dimensional aqueous-derived biomaterial scaffolds from Silk Fibroin. *Biomaterials* **26**(15), 2775–2785 (2005)
17. T.D. Gordon, L. Schloesser, D.E. Humphries, M. Spector, Effects of the degradation rate of collagen matrices on articular chondrocyte proliferation and biosynthesis in vitro. *Tissue Eng.* **10**(7–8), 1287–1295 (2004)
18. K. Hu, F. Cui, Q. Lv, J. Ma, Q. Feng, L. Xu, D. Fan, Preparation of Fibroin/recombinant human-like collagen scaffold to promote fibroblasts compatibility. *J. Biomed. Mater. Res. A* **84**(2), 483–490 (2008)
19. Q. Lv, K. Hu, Q. Feng, F. Cui, Fibroin/collagen hybrid Hydrogels with crosslinking method: preparation, properties, and cytocompatibility. *J. Biomed. Mater. Res. A* **84**(1), 198–207 (2008)

20. P. Chomchalao, S. Pongcharoen, M. Sutteerawattananonda, W. Tiyaboonchai, Fibroin and Fibroin blended three-dimensional scaffolds for rat chondrocyte culture. *Biomed. Eng. Online* **12**, 28 (2013)
21. S.K. Samal, M. Dash, F. Chiellini, X. Wang, E. Chiellini, H.A. Declercq, D.L. Kaplan, Silk/chitosan biohybrid Hydrogels and scaffolds via green technology. *RSC Adv.* **4**, 53547 (2014)
22. S. Thomas, Alginate dressings in surgery and wound management—part 3. *J. Wound Care* **9**(4), 163–166 (2000)
23. K. Ziv, H. Nuhn, Y. Ben-Haim, L.S. Sasportas, P.J. Kempen, T.P. Niedringhaus, M. Hrynyk, R. Sinclair, A.E. Barron, S.S. Gambhir, A tunable Silk-alginate Hydrogel scaffold for stem cell culture and transplantation. *Biomaterials* **35**(12), 3736–3743 (2000)
24. J. Ming, Z. Jiang, P. Wang, S. Bie, B. Zuo, Silk Fibroin/sodium alginate fibrous Hydrogels regulated hydroxyapatite crystal growth. *Mater. Sci. Eng. C* **51**, 287–293 (2000)
25. J. Ming, S. Bie, Z. Jiang, P. Wang, B. Zuo, Novel hydroxyapatite nanorods crystal growth in Silk Fibroin/sodium alginate nanofiber Hydrogel. *Mater. Lett.* **126**, 169–173 (2014)
26. E.S. Gil, D.J. Frankowski, R.J. Spontak, S.M. Hudson, Swelling behavior and morphological evolution of mixed gelatin/Silk Fibroin Hydrogels. *Biomacromolecules* **6**(6), 3079–3087 (2005)
27. X. Hu, D. Kaplan, Silk biomaterials, in *Comprehensive Biomaterials*. 207–19 (2011)
28. J.W. Nichol, S.T. Koshy, H. Bae, C.M. Hwang, S. Yamanlar, A. Khademhosseini, Cell-laden microengineered gelatin methacrylate Hydrogels. *Biomaterials* **31**, 5536–5544 (2010)
29. H. Aubin, J.W. Nichol, C.B. Hutson, H. Bae, A.L. Sieminski, D.M. Cropek, P. Akhyari, A. Khademhosseini, Directed 3D cell alignment and elongation in microengineered Hydrogels. *Biomaterials* **31**(27), 6941–6951 (2010)
30. X. Hu, Q. Lu, L. Sun, P. Cebe, X. Wang, X. Zhang, D.L. Kaplan, Biomaterials from ultrasonication-induced Silk Fibroin-hyaluronic acid Hydrogels. *Biomacromolecules* **11**(11), 3178–3188 (2010)
31. R. Elia, D.R. Newhide, P.D. Pedevillano, G.R. Reiss, M.A. Firpo, E.W. Hsu, D.L. Kaplan, G.D. Prestwich, A. Peattie, Silk-hyaluronan-based composite Hydrogels: a novel, securable vehicle for drug delivery. *J. Biomater. Appl.* **27**(6), 749–762 (2013)
32. M. Pavlovic, X. Serban, N. Yu, M.J. Manesis, Cross-linked Silk-hyaluronic acid compositions. Google Patents, US Patent App. 13/868,010 (2013)
33. D. Zhang, K. Chen, L. Wu, D. Wang, S. Ge, Synthesis and characterization of PVA-HA-Silk composite hydrogel by orthogonal experiment. *J. Bionic Eng.* **9**, 234–242 (2012)
34. T. Suopajarvi, E. Koivuranta, H. Liimatainen, J. Niinimäki, *J. Environ. Chem. Eng.* **2**, 2005–2012 (2014)
35. H.J. Kim, Y.J. Yang, H.J. Oh, S. Kimura, M. Wada, U.-J. Kim, (Springer, 2017). <https://doi.org/10.1007/s10570-017-1491-7>
36. Y. Liu, J. Lim, S.H. Teoh, Review: development of clinically relevant scaffolds for vascularised bone tissue engineering. *Biotechnol. Adv.* **31**(5), 688–705 (2013)
37. M. Fini, A. Motta, P. Torricelli, G. Giavaresi, N. Nicoli Aldini, M. Tschon, R. Giardino, C. Migliaresi, The healing of confined critical size cancellous defects in the presence of Silk Fibroin Hydrogel. *Biomaterials* **26**(17), 3527–3536 (2005)
38. M. Samee, S. Kasugai, H. Kondo, K. Ohya, H. Shimokawa, S. Kuroda, Bone morphogenetic protein-2 (BMP-2) and vascular endothelial growth factor (VEGF) transfection to human periosteal cells enhances osteoblast differentiation and bone formation. *J. Pharmacol. Sci.* **108**(1), 18–31 (2008)
39. M. Liu, X. Zeng, C. Ma, H. Yi, Z. Ali, X. Mou, S. Li, Y. Deng, N. He, Injectable Hydrogels for cartilage and bone tissue engineering. *Bone Res.* **5**, 17014 (2017)
40. F. Mirahmadi, M. Tafazzoli-Shadpour, M.A. Shokrgozar, S. Bonakdar, Enhanced mechanical properties of thermosensitive chitosan Hydrogel by Silk fibers for cartilage tissue engineering. *Mater. Sci. Eng. C* **33**, 4786–4794 (2013)
41. E.G. Lima, L. Bian, K.W. Ng, R.L. Mauck, B.A. Byers, R.S. Tuan, G.A. Ateshian, C.T. Hung, The beneficial effect of delayed compressive loading on tissue-engineered cartilage constructs cultured with TGF-beta3. *Osteoarthritis Cartilage* **15**(9), 1025–1033 (2007)

42. K.D. Kochanek, J. Xu, S.L. Murphy, A.M. Miniño, H.C. Kung, Deaths: final data for 2009. *Natl. Vital Stat. Rep.* **60**(3), 1–116 (2007)
43. M. Floren, W. Bonani, A. Dharmarajan, A. Motta, C. Migliaresi, W. Tan, Human mesenchymal stem cells cultured on Silk Hydrogels with variable stiffness and growth factor differentiate into mature smooth muscle cell phenotype. *Acta Biomater.* **31**, 156–166 (2016)
44. W. Sun, T. Incitti, C. Migliaresi, A. Quattrone, S. Casarosa, A. Motta, Genipin-crosslinked gelatin-Silk Fibroin Hydrogels for modulating the behaviour of pluripotent cells. *J. Tissue Eng. Regen. Med.* **10**(10), 876–887 (2016)
45. C.S. Kim, Y.J. Yang, S.Y. Bahn, H.J. Cha, A bioinspired dual-crosslinked tough Silk protein Hydrogel as a protective biocatalytic matrix for carbon sequestration. *NPG Asia Mater.* **9**, e391 (2017)
46. C. Yan, A. Altunbas, T. Yucel, R.P. Nagarkar, J.P. Schneider, D.J. Pochan, Injectable solid Hydrogel: mechanism of shear-thinning and immediate recovery of injectable  $\beta$ -hairpin peptide Hydrogels. *Soft Matter* **6**(20), 5143–5156 (2010)
47. J.Y. Fang, J.P. Chen, Y.L. Leu, H.Y. Wang, Characterization and evaluation of Silk protein Hydrogels for drug delivery. *Chem. Pharm. Bull. (Tokyo)* **54**(2), 156–162 (2006)
48. Z. Gong, Y. Yang, Q. Ren, X. Chen, Z. Shao, Injectable thixotropic Hydrogel comprising regenerated Silk Fibroin and hydroxypropylcellulose. *Soft Matter* **8**, 2875–2883 (2012)
49. Z. Ding, H. Han, Z. Fan, H. Lu, Y. Sang, Y. Yao, Q. Cheng, Q. Lu, D.L. Kaplan, Nanoscale Silk–Hydroxyapatite Hydrogels for injectable bone biomaterials. *ACS Appl. Mater. Interfaces* 16913–16921 (2017)
50. H.W. Ju, O.J. Lee, B.M. Moon, F.A. Sheikh, J.M. Lee, J.-H. Kim, H.J. Park, D.W. Kim, M.C. Lee, S.H. Kim, C.H. Park, H.R. Lee, Silk Fibroin based Hydrogel for regeneration of burn induced wounds. *Tissue Eng. Regenerative Med.* **11**, 203–210 (2014)

# Long Non-coding RNA for Plants Using Big Data Analytics—A Review



P. Swathi, S. Jyothi, and A. Revathi

**Abstract** This study delves into long non-coding RNAs (lncRNAs). The Long non-coding RNAs (lncRNAs) framework is a noteworthy section in non-coding RNAs and affects numerous biological processes. The long non-coding RNAs (lncRNAs) are not converted code to proteins. They are described as transcripts which have lengths in excess of 200 bp (Nucleotides). Long non-coding RNAs (lncRNAs) are deemed important in the development of plant species as well as their stress responses. All types of lncRNAs are being linked to a variety of developmental processes, diseases, as well as stress in plants. The scientific establishments can benefit from the formation of newer databases and also upgrading of existent databases by providing researchers easy accessibility to lncRNAs' knowledge-base. To achieve this, Big Data analytics in storage of data and Machine learning algorithms for implementation plays a major role. lncRNAs is a prospectively vital class of RNA and has less appropriate prediction tools and databases. With few endorsed researches into lncRNAs of plants, Big data and ML algorithms could be pivotal. Finally, we present the role of emergent systems and databases to store the data of lncRNAs of plants.

**Keywords** Long non-coding RNAs (lncRNAs) · Nucleotides · Big data analytics · Machine learning · Database

## 1 Introduction

Non-coding RNAs with a length of more than 200 nucleotides are termed as Long non-coding RNAs (lncRNAs). They contribute significantly towards gene transcription, epigenomic regulation, and expression of protein-coding genes [1]. In the recent

---

P. Swathi (✉) · S. Jyothi · A. Revathi  
Department of Computer Science, SPMVV, Tirupati, India  
e-mail: [swathivinubaby@gmail.com](mailto:swathivinubaby@gmail.com)

S. Jyothi  
e-mail: [Jyothi.spmvv@gmail.com](mailto:Jyothi.spmvv@gmail.com)

A. Revathi  
e-mail: [arevathi20@gmail.com](mailto:arevathi20@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_21](https://doi.org/10.1007/978-3-030-46939-9_21)

decades, lncRNAs have been getting world-wide attention as a possibly new and critical factor in biological regulation. Many types of lncRNAs are associated to wide a broad spectrum of diseases as well as developmental processes. However, the knowledge of the inner workings of lncRNAs and their functions remain inadequate and it is claimed that almost the whole of the mammalian genome is transcribed into functional non-coding transcripts. In the field lncRNA, at present, action mechanisms of lncRNAs, biological functions, and genomic contexts, are known.

Research into plants, in the recent past, reveal that several long non-coding RNAs (lncRNAs) exist in plants and also can be determined. The role of such long non-coding RNAs (lncRNAs) is to regulate reproductive development [2]. lncRNAs were originally dismissed as pure “noise” in genomes. However, lncRNAs continue to attract interest from researchers around the world. It has also been possible to functionally characterize many lncRNAs at both transcriptional and post-transcriptional levels [3–5].

Massive volume of information is generated when researchers sequence, map, and analyze genomes. This essentially drives genomic studies into the domain of Big Data. Genomic studies produce massive amounts of data; every genome consists of more than 20,000–25,000 genes consisting 3 million base pairs. Thus, it generates more gigabytes data of data comparable to Lakhs of photographs. Sequencing multiple genomes of human or animal or plants may rapidly swell to petabytes of data running into hundreds and even thousands. Data generated from gene-interaction analyses further increases the already colossal amounts of existing data.

## 2 Long Non-coding RNA (LncRNA)

### 2.1 Discovery and Identification of LncRNA

As we know that coding and non-coding RNA's in plants are very important. The recent past however, has seen the emergence of a class which involves RNAs capable of both protein coding as well as non-coding functionality. These ‘coding and non-coding RNAs’ (cncRNAs) are bi-functional. They are instrumental in distinctive developmental mechanisms in plants.

In the period of NGS (Next-generation sequencing), the high-throughput RNA-seq data have been enlighten with the requirement of the non-coding component in a genome in the functions of genes. Previously termed as ‘junk DNA’, ncRNAs are construed from non-coding DNA. A deeper look into transcriptomes from myriad classes revealed that it is possible to transcribe around 90% of a genome; wherein just a tiny fraction of transcribed regions potentially codes for proteins [6]. Categorization of ncRNAs is done based on their expressions and functions in various types of cells; and can be classified into organizing and regulatory ncRNAs. t-RNA, snRNA and r-RNA are examples of housekeeping ncRNAs. The expressions of these are prominent and have a structural function in all cells [7]. On the other hand, temporal expressions

are exhibited in particular types of cells by regulatory ncRNA. These include small interfering RNAs (siRNAs), promoter-associated RNAs (PARs), long non-coding (lncRNA), microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), and enhancer RNAs (eRNAs). To identify lncRNAs amongst all organisms, there is a set yardstick of >200 nt length [8]. In ncRNAs, a large portion is taken up by lncRNA. It uses various molecular mechanisms for regulating a variety of biological processes.

As of now, detailed characterization of only a few lncRNAs has been done; all lncRNAs have not been characterized. Nonetheless, researchers know that lncRNAs are important and that they regulate gene expressions. It is also known that the vast range of functions of lncRNAs affect cellular and developmental processes. With the use of a variety of diverse mechanisms, lncRNAs can perform both gene activation and gene inhibition. This adds an additional layer of complexities while trying to understand genomic regulation. Estimation has put that between 25 and 40% of coding genes contain overlapping antisense transcription. Hence, one should not underestimate the effects of lncRNAs on gene regulation.

The domain of bioinformatics is facing a continuous challenge in accurately identifying and in functional annotating high-throughput RNA-seq data. The lncRNAs in plants are identified and this data is transferred and added into various databases [9]. High-confidence lncRNAs are assembled and identified through a pipeline designed with multiple filters, as shown in Fig. 1.

## 2.2 Characteristics and Prominent Features of LncRNAs

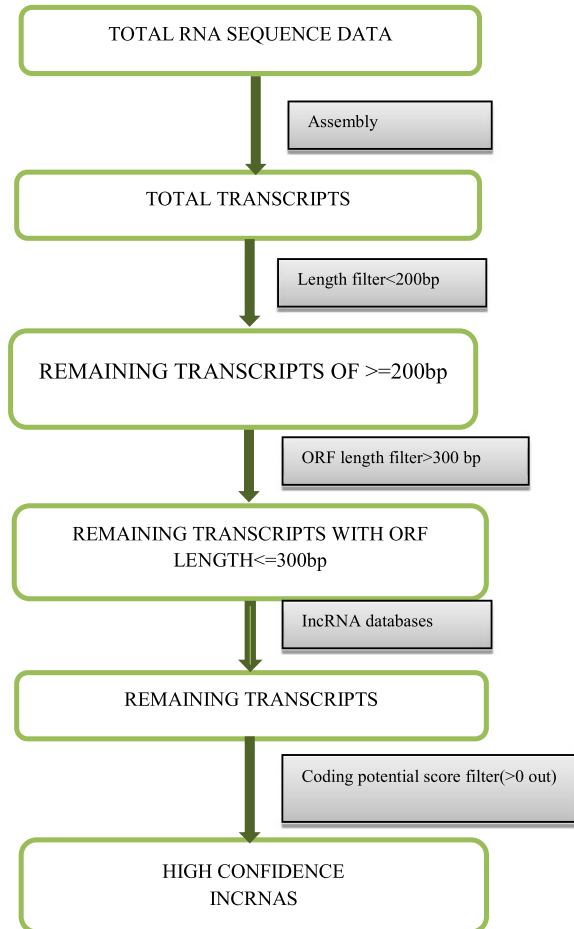
Genomic locations of lncRNAs biotypes, with respect to neighbouring genes, can be used to define lncRNAs. They can be broadly categorized into: (1) intergenic, is the lncRNA which occurs at genomic intervals between 2 genes; (2) intronic, is inferred from associate degree intron; (3) The sense, once it finishes overlapping the exons on constant strands; (4) The antisense, the exons on opposite strands are overlapped; (5) The bidirectional, when the lncRNA were expressional and an head-to-head committal to write transcript on alternative strand is initiated in shut genomic proximity [10].

There is a similarity with mRNAs in several ways; though generally akin to lncRNAs, they do not have the ability of protein-coding [11] like RNA polymerases will take the responsible for lncRNA transcription, polyadenylation, 5' capping, and alternative splicing patterns [12]. The majority of lncRNAs are spliced with exon/intron lengths similar to that of mRNA coding genes [13, 14].

Recent studies in Tissue similarity were conducted. These researches have indicated that tissue-specific lncRNAs started at a specific stage in the development cycle. Conclusions of studies indicate that lncRNAs are incidental to development of fibers in *G. arboretum* [15], development of flowers in *C. arietinum* [16], development of flowers and fruits in *Fragaria vesca* (woodland strawberry) [17], and development of floral organs and roots in *Morus notabilis* (mulberry) [18].



**Fig. 1** Pipeline to identify long non-coding RNA [40]



### 2.3 Emerging Significance of lncRNAs and Status in Plants

It is proven that over decades, the evolution studies of lncRNA in the plants are behind when compared with research in mammals, including humans. Though at the outset, it was prompted by full-length cDNA cloning and tiling microarrays, detection of plant lncRNAs is currently more concentrated and encouraged by cutting edge RNA-seq.

Current research has discovered plant lncRNA genes namely AtIPS1 [19], COOLAIR [20], COLDAIR [2], LDMAR [21, 22]. On the other hand, Older research revealed plant lncRNA genes, which are considered in biologically vital, namely GmENOD40 [23], MtENOD40 [24], TPS11 [25], and OsPII [26]. These studies have presented insights into the varied roles that such long RNA play biologically. The lncRNAs in plants determine numerous biological processes and molecular

functions in them; such as fertility, vernalization, photo morphogenesis, phosphate homeostasis, protein re-localization, modulation of chromatin loop dynamics, alternative splicing, etc. In these lncRNA genes, a significant amount of data has been gathered due to the recent prevalent applications of high-throughput RNA-seq and computational pipelines. Expansion in the volume of such RNA-seq data is now available publically. This facilitates global identification and silico characterizing of lncRNAs in a variety of species of plants; and also laying foundation for researchers to understand the potential functions and expression patterns of lncRNAs.

## ***2.4 Latest Advances in Studying Plant LncRNAs***

Researchers have focused their studies on lncRNAs in mammals thus far. However, identification of lncRNAs in plants has nearly caught up with studies in lncRNAs in mammals. Databases in which lncRNAs of plants could be accessed are listed in Table 1.

### **2.4.1 Why Big Data in Genomics Now?**

Studies show that using technologies such as Big Data and analytics could save billions of dollars in genomic research. Big Data aids in improving operational efficiency and also in predicting, planning disease and epidemic responses and enhances the quality of clinical trials monitoring. It also optimizes expenditure on healthcare by patients, by hospitals and government organizations. Genomic sequencing is a vital sector and it is envisaged to be the future of lncRNA. We consider the potential of technologies such as Big Data and analytics to uncover the role lncRNA plays in plants (Fig. 2).

### **2.4.2 Machine Learning for Genomics**

In recent times, ensemble methods have gained popularity in order to solve complex biological problems generally resolved using machine learning. A single model can be found from multiple learners by working with ensemble models. This overcomes issues such as over-fitting; and it supports generalization of the classifier. Along with improvements in classification, collective approaches do away with the difficult choice of which is the “best” model. This is because a single classifier is capable of enveloping all models. Every distinct classifier, which is incorporated in creating the whole ensemble model, has its own strengths in terms of classification. This results in predictions which are robust and precise when a combination of these classifiers is used in conjunction [27].

The objective of several machine learning problems is to discover a single model that will best expected outcome that we want. Other than making one new model and

trusting this new model is the best/most accurate predictor we can make, ensemble methods take numerous models into interpretation, and finds the average of those models to produce one final and best model. We can't say that Decision Trees are not the only form of ensemble methods, just the most popular and relevant in Data Science today which is help full for predictions in genomic world also.

**Table 1** Databases of plant lncRNAs

Database	Descriptions/features	Website
TAIR, The Arabidopsis Information Resource	The database includes the comprehensive which includes genomic sequences, structure gene and coding and non-coding proteins. Mapping tools also available	<a href="http://www.arabidopsis.org">www.arabidopsis.org</a>
Araportll	The databases which are based on Araport11 and also includes coding and non-coding. These are additional and can be compared with TAIR10.lincRNAs, ncRNAs and NTA compared with TAIR10	<a href="http://www.araport.org">www.araport.org</a>
PLncDB (Plant long noncoding RNA database)	The database contains more than 13 thousand long non-coding RNAs which are identified using RNA-seq. PLncDB was used for genome browser which is associated with several epigenetic markers	<a href="http://www.chulab.rockefeller.edu">www.chulab.rockefeller.edu</a>
GREENC (Green Non-coding database)	Green non coding Database have more than 12,000 different lncRNAs from 37 different species of plants. The user has the access for folding energy and potential for each and every Long non coding RNAs	<a href="http://www.greenc.sciencedesigners.com">www.greenc.sciencedesigners.com</a>
NONCODE v4.0	There are more than 500,000 NON CODE Long non coding RNAs are available from 16 multi species. The only plant was Arabidopsis, as non-coding which is focused on plants species, rest are all belongs to Humans and Mouse	<a href="http://www.noncode.org">www.noncode.org</a>

(continued)

**Table 1** (continued)

Database	Descriptions/features	Website
CANTATAdb	More than 45,000 plants lncRNAs are available in CANTATAdb from 10 different species. The tissue specific expression and potential coding for each RNA is also calculated based on the splicing potential role and miRNA modulation	<a href="http://www.cantata.amu.edu">www.cantata.amu.edu</a>
PNRD (Plant ncRNA database)	In this database the analytics tools like miRNA predictor, customized browser for genome. The 11 types of human and 150 plants are included which makes more than 25 thousand ncRNAs available	<a href="http://www.structuralbiology.cau.edu.cn">www.structuralbiology.cau.edu.cn</a>
PlantNATsDB (Plant Natural Antisense Transcripts DataBase)	All natural antisense transcripts in this database which are from different plants species are 70	<a href="http://www.bis.zju.edu.cn">www.bis.zju.edu.cn</a>

Bio-sequencing technologies have its limitations. Along with such limitations, the process of sequencing itself is prone to errors; some of which are unavoidable. Owing to these factors, the recent years has witnessed development in predictive tools for mRNAs and lncRNAs. The similarity amongst such developed tools is the implementation of machine learning. These tools use machine learning for training recognition models of mRNAs and lncRNAs. Deep Learning is a novel machine learning technique. This has been employed in an assortment of biological studies including structural biology, transcriptomics, genomics, and proteomics. However, only marginal researches have implemented and used deep learning networks to identify lncRNAs.

In machine learning, there are two major paradigms: unsupervised learning and supervised. Both of these can be applied in the field of biology.

In supervised learning, according to set of attributes or feature the collection of objects are taken. The results of the classification method may be a set of rules which cause assigning of objects to categories based mostly only on values of options. In the context of biology, samples of mapping of objects to classes, square measure tissue organic phenomenon profiles to sickness cluster, and super molecule sequences to their secondary structures.

Unsupervised learning, contrary to supervised learning, does not have predefined class labels for objects being studied. The aim here is data exploration and discovering similarities among objects. Similarities square measure won't outline teams of objects, noted as clusters.



**Fig. 2** Big data in genomics [41]

Unsupervised learning intends to uncover naturally occurring groupings in data. The contrast between unsupervised and supervised learning can be noted as: All data-inputs, in unsupervised learning, are unlabeled, and the learning method does both define labels and associate objects to them; all data-inputs, in supervised learning, consists of class labels, learning is focused on associating labeled-data to classes.

### 2.4.3 Machine Learning Based Tools

The algorithms are designed such that they compute and infer functions from sample data. Simply put, machine learning is capable of learning from data and adapting to any progress and changes; and it does not need specific coding and programming [28].

There are some machine learning based for lncRNA identification in plants and multi-species. Different machine learning algorithms are used by different tools for constructing classifiers. The various tools available, which employ support vector machine (SVM), are CNCI, lncRScan-SVM, CPC, CPC2, and PLEK. Logistic regression is used by CPAT and lncScore tools. Random forest or balanced random

**Table 2** Provides a summary of lncRNA identification tools based on machine learning

Tool names	Year	Input format	Species	Requirement	Model	References
CPC	2007	FASTA	Multi-Species	Linux BLAST	SVM	[29]
CNCI	2013	FASTA GTF	Multi-Species	Linux Python 2.7	SVM	[30]
CPAT	2013	FASTA BED	Multi-Species	Linux Python 2.7	Logistic Regression	[31]
PLEK	2014	FASTA	Plant, Vertebrate	Linux Python 2.7	SVM	[32]
COME	2016	GTF	Plant, Worm, Fly	unknown	Balanced Random forest	[33]
CPC2	2017	FASTA GTF	Multi-Species	Linux Python	SVM	[34]
PLncPRO	2017	FASTA	Plants	BLAST Python 2.7	Unknown	[35]
RNAplonc	2019	FASTA	Plants	PERL Python 2.7 R	REPTree	[36]
PLIT	2019	FASTA	Plants	unknown	Random Forest Model	[37]

forest is employed by COME, lncRNA-ID, and FEELnc tools [46, 47]. Deep learning algorithm and deep neural network (DNN) is used to construct DeepLNC (Table 2).

**Coding Potential Calculator**—Coding Potential Calculator is familiarly known as CPC. Transcriptomes studies have shown that a vast quantity of mammalian transcripts does not encode proteins; instead, these act as non-coding RNAs (ncRNAs). This also holds true for other organisms. Vivo experimentations have verified that non-coding RNAs play vital biological roles of regulating transcriptions and translation. There is a lot of interest shown, in the biological research circles, towards studying and identifying newer non-coding RNAs.

To measure the potential for coding of a transcript, six features were extracted from the nucleotide sequence of a transcript. In comparison to a non-coding transcript, a correct protein-coding transcript has a strong possibility of containing a long and high-quality Open Reading Frame (ORF). Therefore, the extent and quality of a transcript's ORF is measured by the first three, out of the six, extracted features. The software 'framefinder' (32) was used for identification of the longest-reading frame out of three forward-frames. The enormous protein sequence databases are expanding rapidly. These databases offer extensive information for identifying protein-coding transcripts.

The extracted six features acted as inputs for a widely used learning classifier named support vector machine (SVM). First, input features are mapped to a high-dimensional feature space using appropriate kernel functions. Then, the SVM assembles a classification hyper-plane (maximum margin hyper-plane) for separating the converted data-points. SVM has a wide acceptance amongst researchers for its stable performance and high precision. Thus, SVM is a classification tool which is extensively employed for analysis in bioinformatics. These applications include cancer classifications based on microarrays, protein-function predictions, and sub-cellular localization predictions.

**Coding-Non-coding Index**—A robust signature tool was constructed and assessed in this study, termed as Coding-Non Coding Index (CNCI). For this, we profiled adjoining nucleotide triplets (ANT) to efficiently differentiate sequences which are non-coding and protein-coding unrelated to existent annotations. CNCI is efficient in classification of incomplete transcripts and sense–antisense pairs. Application of CNCI presented precise classifications of transcripts compiled out of data containing whole-transcriptome sequencing across species. This offered insights into divergence of evolutionary genes between invertebrates and vertebrates, and between plant species; it also offered a directory of long non-coding RNAs in *Pongo pygmaeus*.

This research analyzed the usage-frequency of ANT in coding domain sequence (CDS) and sequences of ncRNA by implementing a sliding window. The classification was used to find out frequency in ANT. An important step for this method is identifying the CDS region of all transcripts. To find the size of sliding window for achieving the best final classification result, N number of series with different lengths has been found; human data was used to train a classifier.

By using prediction and classification five features were extracted i.e. the codon-bias, based on the length and MLCDS' S-score, the length percent, and distance score. CNCI has couple of steps, includes scoring the sequence and classification model construction.

**Coding Potential Assessment Tool (CPAT)**—It is a new alignment-free methodology. CPAT, robustly identifies non-coding and coding transcripts out of an enormous set of potential candidates. A logistic regression model is used by CPAT. These are built with four sequence features namely, Fickett TESTCODE statistic, open reading frame coverage, hexamer usage bias, and open reading frame size. CPAT is highly accurate. CPAT is virtually four times faster than Phylo Codon Substitution Frequencies and Coding Potential Calculator. This empowers processing more than a thousand transcripts in just a few seconds. It reads input-sequence files which are formatted as FASTA- or BED- formats. A web-based interface of CPAT, permits submission of sequences and retrieval of prediction results almost instantaneously.

Logistic regression a suited methodology; this is due to predicting the potential for coding primarily being a binary decision problem. It is an alignment-free methodology; thus, the calculation of all selected features (predictor variables) is done from the sequence directly. The max-length of Open reading frame (ORF) was the first feature. ORF coverage was the second feature; it is the ORF to transcript length ratio. The third feature used was the Fickett TESTCODE score. And, hexamer usage

bias was the fourth feature. These four features were used as predictor variables in order to construct a logistic regression framework. To evaluate the performance of the prediction of the logic model, a high-confidence training dataset was constructed. C and Python were used to implement the source code; the code is openly accessible. PHP, MYSQL and Apache, was applied on the web server. Support for Major web-browsers was also provided for.

**Predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme**—The performance of PLEK has been assessed on well-annotated transcripts of lncRNA and mRNA. A tool named PLEK, which is a characteristic k-mer based alignment-free tool, was developed in this study. This tool was constructed to distinguish mRNAs from lncRNAs. In order to accomplish this, the computational features which PLEK takes as inputs are the calibrated k-mer frequencies of a transcript sequence. Then, a classification model was built to distinguish lncRNAs from mRNAs. This classification model was binary in nature and was built using the support vector machine (SVM) algorithm. PLEK's performance other vertebrate data was good; it used classification models trained from human datasets.

After collecting the data from different sources, an SVM-scale program has been implemented from LIBSVM package. Then, a support vector machine (SVM) with a radial basis functional kernel, whose variance is gamma, was selected as the binary classifier. The LIBSVM package has a script named grid.py. This script was used to obtain the gamma of the kernel and the optimal C of the SVM. It was found that PLEK performed well. This was after testing it many different vertebrate datasets to compute its performance in cross-species prediction of protein coding and non-coding transcripts.

**A coding potential calculation tool based on multiple features**—To recognize and characterize newer lncRNAs from new sample data of RNA sequencing, a tool named COME was developed based on many features. It incorporates many features derived from sequences and based on experiments with the use of decompose–compose approach. This makes this tool's accuracy and robustness better than other widely used tools [37]. The tool COME also showed consistent and substantial improvement in predication results over existing coding potential calculators. Unlike other tools, this tool also explains and characterizes all predicted transcripts of lncRNAs with several lines of evidence which supports it. By using the well-validated database knowingly lncRNAdb, and classified supporting features a subgroup of lncRNAs were found.

To compute the score of the coding potential of a transcript, COME implemented machine learning models by integrating several features obtained from experimental data and sequence data. Based on the composed matrix, we applied a balanced random forest (BRF) algorithm to train on the annotated coding (mRNAs) and non-coding transcripts (lncRNAs) [37]. A five-fold cross validation on was applied on training data for model optimization in COME. For the unbalanced training set the BRF algorithm which used multiple sub-training sets, the same number of mRNAs and lncRNAs was used. The decompose–compose method and multiple features enabled COME to compute the coding potential accurately, robustly and consistently [37].



**Coding potential calculator 2**—In order to estimate the coding capability of RNA transcripts accurately and fast, the coding potential calculator has been upgraded to CPC2 from CPC1. CPC2 is 1000 times quicker in comparison to CPC1. It is also more accurate than CPC1; specifically for lncRNA transcripts. Furthermore, the CPC2 model is neutral to various species. This makes it conceivable for constantly increasing transcriptomes of non-modeled organisms. To assess CPC2's performance with different species, separate testing sets were built for humans, mice, zebrafish, flies, worms and the Arabidopsis plant.

A support vector machine (SVM) model was trained with the use of four essential features. In order to train the SVM model, an LIBSVM package [38] was implemented with the use of a standard RBF kernel (radial basis function kernel) along-with the training database consisting of 17,984 high-confidence Homo sapiens protein-coding and 10,452 non-coding transcripts [31].

**Plant Long Non-Coding RNA Prediction by Random forest**—Perfect annotation and identification of lncRNAs is the most important stage to get in-depth understanding with regards to their functions. PLncPRO is a unique tool which predicts plant lncRNAs with the use of transcriptome data. PLncPRO is ideal for plant species, and has higher predicting accuracy in comparison with existent tools [39]. To help in predicting lncRNAs in non-model/orphan plants some consensus models for dicots and monocots were developed. Investigation of differential expression and various characteristics under the salinity and stressful conditions was done, and lncRNAs were validated via RTqPCR [39].

In addition, the application of the tool predicted newer lncRNAs in two crops, rice and chickpea, under abiotic stress conditions. They used 2 programs, firstly Framefinder for estimating the quality of an open reading frame (ORF) existent in a transcript, and secondly the BLASTX was used to extract when the transcripts has a substantial resemblance, protein coding sequence to known by any and used SWISS-PROT database [39].

For implementation of PLncPRO, Random Forest Model was used. It is an ensemble learning technique for regression and classification. It builds many decision trees from a training dataset and outputs the computed result in the forest for all the trees available. The large datasets are very effectively handled by the random forest with many variables which makes the ranking as relative feature, this helps the model and data to interpretation of them. Cross-validation is not needed for RFM, because it uses the report out of bag model which is very similar to the cross-validation.

**RNAplonc**—There exist inadequate explicit computational methodologies for reliable prediction of lncRNAs in plants, in comparison to an array of available prediction tools for lncRNAs in mammals. RNAplonc predicts lncRNA from plant transcripts with greater reliability; the best result of 87.5% from eight different tests and species which are taken from the database called GreenNC.

The RNAplonc has provided Four main new contributions: (i) a comprehensive analyses of 5468 features along with the different three approaches for ten machine learning algorithm were used for feature-selection (ii) 16 detected features to vigorously differentiate lncRNAs and mRNAs in plants; (iii) in juxtaposition with other widely used tools, RNAplonc accurately and robustly detects lncRNAs gathered

from various databases and case-study datasets; and (iv) a framework to ensure RNAplonc remains a user-friendly tool, capable of identifying lncRNAs while using lesser computational resources [46].

In the RNAplonc tool, the implementation was done by using machine learning approach and a classifier to identify lncRNAs in plants. The model was built using tuning parameters and cross-validation techniques. The analysing was done by using 8 different MLA (J48, Naive Bayes, IBK, Random Forest, Multilayer Perceptron, Random Tree, SVM and REPTree). Tuning parameter techniques were applied to find the best parameters for each model; and then 10-fold cross-validation techniques were used by each trained model. The methods had similar performance, as the cluster of features was distinctive. For lesser computational prediction cost the standard classifier for PNAplonc, REPTree was chosen.

**Plant LncRNA and identification Tool (PLIT)**—PLIT implements a feature choice methodology which supports L1 regularization and iterative Random Forests (iRF) categorization. This yields choices of options which are optimum. It categorizes RNA-seq derived FASTA sequences into writing or long non-coding transcripts. This is done on the basis of sequence and codon-bias options. Upon L1 regularization, thirty one optimum options were obtained supported lncRNA and protein-coding transcripts among eight species of plants. The tool's efficiency had been assessed using 10-fold cross-validation on the datasets of seven plants' RNA-seq.

The investigation indicates greater veracity when in contrast to presently available progressive CPC tools. A threshold of 200 bp on non-coding RNA (ncRNA) FASTA files was put in place. This threshold was applied to filter lncRNA sequences. The architecture incorporates an optimization module named LiRFFS. LiRFFS stands for LASSO iterative Random Forest-Feature Selection. It chooses an optimum feature-set. This feature-set is selected from the features of training and validation data-set. The feature-set which is selected is utilized to identify lncRNAs straight out of RNA-seq derived FASTA sequences.

### 3 Conclusion

In this review, we have highlighted numerous approaches pertaining to long non-coding RNAs in plant species. Genomes in plants encode a vast quantity of non-coding RNAs, chiefly lncRNAs, which concern stress tolerance and disease control. Big data analytics has been used to store huge amounts of data. Owing to the accelerated pace at which genomic data is accumulating, we have used machine learning algorithms for developing unique and integrated tools for analyzing the functions of lncRNAs. Resultant evidence could offer newer biological perspectives about the regulatory roles they perform in plants. Most of the recent studies only exploit the power of big data; however, subsequent endeavors could see implementation of distributed computing along with big data for better results.

## References

1. S. Swiezewski, P. Crevillen, F. Liu, J.R. Ecker, A. Jerzmanowski, C. Dean, SmallRNA-mediated chromatin silencing directed to the 3' region of the Arabidopsis gene encoding the developmental regulator, FLC. *Proc. Natl. Acad. Sci. USA* **104**, 3633–3638 (2007)
2. J.B. Heo, S. Sung, Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331**, 76–79 (2011)
3. B.B. Amor, S. Wirth, F. Merchan, P. Laporte, Y. D'Aubenton-Carafa, J. Hirsch, A. Maizel, A. Mallory, A. Lucas, J.M. Deragon, Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. *Genome Res.* **19**, 57–69 (2009)
4. Y. Bai, X. Dai, A.P. Harrison, M. Chen, RNA regulatory networks in animals and plants: a long noncoding RNA perspective. *Brief. Funct. Genom.* **14**, 91–101 (2015)
5. J. Zhang, X.W. Cui, Y.H. Shen, L.X. Pang, A.Q. Zhang, Z.Y. Fu, J.T. Chen, X.R. Guo, W.H. Gan, C.B. Ji, Distinct expression profiles of LncRNAs between brown adipose tissue and skeletal muscle. *Biochem. Biophys. Res. Commun.* **443**, 1028–1034 (2014)
6. E.D. Kim, S. Sung, Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci.* **17**, 16–21 (2012)
7. J. Liu, C. Jung, J. Xu, H. Wang, S. Deng et al., Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* **24**, 4333–4345 (2012)
8. J. Liu, H. Wang, N.H. Chua, Long noncoding RNA transcriptome of plants. *Plant Biotechnol. J.* **13**, 319–328 (2015)
9. G. Bhatia, N. Goyal, S. Sharma, S.K. Upadhyay, K. Singh, Present scenario of long noncoding RNAs in plants. *Non Coding RNA* **3**, 16 (2017)
10. C.P. Ponting, P.L. Oliver, W. Reik, Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641 (2009)
11. A.R. Karapetyan, C. Buiting, R.A. Kuiper, M.W. Coolen, Regulatory roles for long ncRNA and mRNA. *Cancers* **5**, 462–490 (2013)
12. T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D.G. Knowles et al., The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012)
13. M.B. Gerstein, C. Bruce, J.S. Rozowsky, D. Zheng, J. Du, J.O. Korbil, O. Emanuelsson, Z.D. Zhang, S. Weissman, M. Snyder, What is a gene, post-ENCODE?. History and updated definition. *Genome Res.* **17**, 669–681 (2007)
14. M. Guttman, I. Amit, M. Garber, C. French, M.F. Lin, D. Feldser, M. Huarte, O. Zuk, B.W. Carey, J.P. Cassady et al., Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009)
15. C. Zou, Q. Wang, C. Lu, W. Yang, Y. Zhang, H. Cheng, X. Feng, M.A. Prosper, G. Song, Transcriptome analysis reveals long noncoding RNAs involved in fiber development in cotton (*Gossypium arboreum*). *Sci. China Life Sci.* **59**, 164–171 (2016)
16. N. Khemka, V.K. Singh, R. Garg, M. Jain, Genome-wide analysis of long intergenic non-coding RNAs in chickpea and their potential role in flower development. *Sci. Rep.* **6**, 33297 (2016)
17. C. Kang, Z. Liu, Global identification and analysis of long non-coding RNAs in diploid strawberry *Fragaria vesca* during flower and fruit development. *BMC Genomics* **16**, 815 (2015)
18. X. Song, L. Sun, H. Luo, Q. Ma, Y. Zhao, D. Pei, Genome-wide identification and characterization of long non-coding RNAs from Mulberry (*Morus notabilis*) RNA-seq Data. *Genes* **7**, 11 (2016)
19. J.M. Franco-Zorrilla, A. Valli, M. Todesco, I. Mateos, M.I. Puga, I. Rubio-Somoza, A. Leyva, D. Weigel, J.A. García, J. Paz-Ares, Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genet.* **39**, 1033–1037 (2007)
20. S. Swiezewski, F. Liu, A. Magusin, C. Dean, Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* **462**, 799–802 (2009)
21. J. Ding, Q. Lu, Y. Ouyang, H. Mao, P. Zhang, J. Yao, C. Xu, X. Li, J. Xiao, Q. Zhang, A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proc. Natl. Acad. Sci. USA* **109**, 2654–2659 (2012)

22. H. Zhou, Q. Liu, J. Li, D. Jiang, L. Zhou, P. Wu, S. Lu, F. Li, L. Zhu, Z. Liu et al., Photoperiod- and thermo-sensitive genic male sterility in rice are caused by a point mutation in a novel noncoding RNA that produces a small RNA. *Cell Res* **22**, 649–660 (2012)
23. W.C. Yang, P. Katinakis, P. Hendriks, A. Smolders, F. Vries, J. Spee, A. Kammen, T. Bisseling, H. Franssen, Characterization of GmENOD40, a gene showing novel patterns of cell-specific expression during soybean nodule development. *Plant J* **3**, 573–585 (1993)
24. M.D. Crespi, E. Jurkevitch, M. Poiret, Y. d'Aubenton-Carafa, G. Petrovics, E. Kondorosi, A. Kondorosi, enod40, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. *EMBO J* **13**, 5099 (1994)
25. C. Liu, U.S. Muchhal, K.G. Raghothama, Differential expression of TPS11, a phosphate starvation-induced gene in tomato. *Plant Mol. Biol.* **33**, 867–874 (1997)
26. J. Wasaki, R. Yonetani, T. Shinano, M. Kai, M. Osaki, Expression of the OsPI1 gene, cloned from rice roots using cDNA microarray, rapidly responds to phosphorus status. *New Phytol.* **158**, 239–248 (2003)
27. C.M.A. Simopoulos, A.E. Weretilnyk, G.B. Golding, *Prediction of Plant lncRNA by Ensemble Machine Learning Classifiers*
28. R.S. Michalski, J.G. Carbonell, T.M. Mitchell, *Machine Learning—An Artificial Intelligence Approach*, vol. I, 1st ed. (Amsterdam, The Netherlands, Elsevier, 1983), pp. 5–6, ISBN 9780934613095
29. L. Kong, Y. Zhang, Z.Q. Ye et al., CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**(Suppl 2), W345–W349 (2007)
30. L. Sun, H. Luo, D. Bu et al., Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **41**(17), e166 (2013)
31. L. Wang, H.J. Park, S. Dasari et al., CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**(6), e74 (2013)
32. A. Li, J. Zhang, Z. Zhou et al., PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improvedk-merscheme. *BMC Bioinformatics* **15**, 311 (2014)
33. L. Hu, Z. Xu, B. Hu, Z.J. Lu, COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res.* **45**(1), e2 (2017)
34. Y.J. Kang, D.C. Yang, L. Kong et al., CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**(W1), W12–W16 (2017)
35. U. Singh, N. Khemka, M.S. Rajkumar, R. Garg, M. Jain, PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Res.* **45**(22), e183 (2017)
36. T.D. Negri, W.A. Alves, P.H. Bugatti, P.T. Saito, D.S. Domingues, A.R. Paschoal, Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants. *Briefings Bioinf.* **20**(2), 682–689 (2019)
37. S. Deshpande, J. Shuttleworth, J. Yang, S. Taramonli, M. England, PLIT: an alignment-free computational tool for identification of long non-coding RNAs in plant transcriptomic datasets. [arXiv:1902.05064v1](https://arxiv.org/abs/1902.05064v1) [q-bio.GN], 12 Feb (2019)
38. C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 27 (2011)
39. C. O'Donovan, M.J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, R. Apweiler et al., High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3**, 275–284 (2002)
40. S. Tyagi, A. Sharma, S.K. Upadhyay, Role of next-generation RNA-seq data in discovery and characterization of long non-coding RNA in plants. Chapter 6, Licensee InTech (2018)
41. Genomics and the role of big data in personalizing the healthcare experience, Bonnie Feldman. O'REILLY. August 22, (2013).

# In Silico, In Vitro and In Vivo Anti-inflammatory and Analgesic Activity of Usnic Acid



D. Sujatha, Ch. Hepsiba Rani, Shaheen Begum, Sunitha Sampathi,  
and Saurabh Shah

**Abstract** This study documents the use of Usnic acid as a prime candidate with a view to predict its analgesic and anti-inflammatory activity, via a mechanistic approach through in silico docking techniques and in vitro assays to extrapolate these results to its in vivo activity. Swiss docking software was used to dock usnic acid with COX, iNOS, MAGL, TRPV1, GABA, glutamate and monoamine oxidase to determine its mechanism. In silico studies proved that usnic acid had similar binding energies compared to standard drug. In vitro assays including HRBC membrane stabilizing assay, albumin denaturation inhibitory assay, proteinase inhibitory assay and lipoxigenase assay further supported in silico study results which was then confirmed by in vivo assays. In silico as well as in vitro data together proved to be fruitful in predicting good analgesic as well as anti-inflammatory activity which can help in reducing animal studies.

**Keywords** Analgesic · Anti-inflammatory · In silico docking studies · Usnic acid

## 1 Introduction

Current scenario of drug research focuses on how to reduce usage of animals for demonstration of their efficacy and safety. Ethical considerations and strict rules and regulations for maintenance and handling of animals, in biomedical research is the major hurdle of in vivo experimentation [1, 2]. This has brought research to a stage where the traditional in vivo animal research needs to be replaced by in silico and in vitro techniques which not only reduces animal experimentation, but also saves the time and cost of the study by predicting and quantifying the activity of drugs in advance which can then be used for dose escalation in humans by various PBPK modeling approach [3, 4]. High throughput screening of a large number of drugs is

---

D. Sujatha (✉) · Ch. Hepsiba Rani · S. Begum  
Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [drsujathasai@gmail.com](mailto:drsujathasai@gmail.com); [drsujathasai@spmvv.ac.in](mailto:drsujathasai@spmvv.ac.in)

S. Sampathi · S. Shah  
Department of Pharmaceutics, NIPER-Hyderabad, Balanagar, Hyderabad, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_22](https://doi.org/10.1007/978-3-030-46939-9_22)

possible within no time via computational studies which can be confirmed by in vitro assays thus minimizing the need of animal experimentation [5, 6]. It exemplifies need of correlation between in silico versus in vitro versus in vivo studies for any selected disease conditions.

Inflammation is a natural and first physiological response of a living tissue to an injury which ultimately aims to limit the damage and promotes tissue repair. Inflammation and many non-communicable diseases like cancer, diabetes, cardiovascular diseases, arthritis were inextricably related. The relationship between inflammation and pain were mediated by cyclooxygenase (COX) enzymes especially with COX2 which help in the synthesis of prostaglandins (PGs) like PGE2 and PGF2a in high concentrations at the inflammatory site [7]. These released PGs either stimulate pain receptor or sensitize pain receptors to the action of other pain producing substances such as histamine, 5-hydroxytryptamine (5HT) and bradykinin which send nociceptive signals to the brain.

Currently, many anti-inflammatory drugs are used in the treatment of acute inflammatory disorders, despite of their renal and gastric adverse effects [8]. These presenting challenges have impelled researchers all over the world to search for safe alternative therapy. According to studies, metabolites of lichens were reported to have various biological activities taken in the form of medications, food supplements and dyes. Usnic acid (UA) is one of the common lichen metabolites, reported to be antibiotic, antiviral, antiprotozoal, larvicidal, antioxidant and have ability to offer UV protection [9]. Although enough scientific information is present on UA, the analgesic activity of it has not been established. Hence, to screen and confirm the efficacy and safety of UA as analgesic, the current work was planned using in silico, in vitro and in vivo approaches. This may also establish UA mechanisms of action for the proposed anti-inflammatory and analgesic activity.

## **2 Materials and Methods**

### ***2.1 Test Drugs and Chemicals***

UA acid was procured from Sigma Aldrich and Ibuprofen was obtained as a gift sample from Alpex International Pvt Ltd. Hyderabad, Telangana, India. The other chemicals and reagents used were of analytical grade and purchased from SD-fine chemicals, Merck India Ltd., Mumbai and Sigma Aldrich.

### ***2.2 Docking Experimental Studies***

Molecular docking using Swiss Dock Molecular Docking calculations were performed using Swiss Dock (based on the docking software EADock DSS). The “grid

(Box size:  $40 \times 40 \times 40$  Å and box center:  $0.38 \times 2.98 \times 20.51$  for x, y and z, respectively) was designed in which many binding modes were generated for the most favorable bindings. Simultaneously, their CHARMM forcefield energies calculated on the grid. Each docking experiment was derived from 250 different consecutive runs. The binding modes with the most favorable energies were evaluated with Fast analytical continuum treatment of solvation (FACTS) and clusters. Binding modes were scored using their Full Fitness and clustered. Clusters were then ranked according to the average Full Fitness of their elements” as reported earlier.

In the studies, “Full Fitness and Gibbs free energy ( $\Delta G$ ) of each run (250 runs) of the docking were evaluated. Favorable binding modes were scored based on Full Fitness and cluster formation. Ranking of the cluster was performed using the value of Full Fitness” as mentioned by the software. Table 1 depicts results obtained from the docking of the into Cyclooxygenase (COX<sub>2</sub>, 3LN1); inducible Nitric Oxide Synthase (iNOS, 2Y37); Mono Acyl Glycerol Lipase (MAGL, 3PE6); Transient Receptor Potential Vanilloid1 (TRPV1, 2PNN); Glutamate (3RN8); I2 (2Z5X) and GABA<sub>A</sub> (Gamma Amino Butyric Acid A, 4COF). Figure 1 shows the hypothetical binding modes of with selected target proteins.

### 3 In Vitro Anti-inflammatory Assays

The anti-inflammatory activity of UA was assessed by using in vitro models like “Human Red Blood Corpuscles (HRBC) membrane stabilizing assay, albumin denaturation, proteinase inhibitory activity and anti-lipoxygenase assay” by following the earlier reported methods [10–13]. The experiments were performed in triplicates. The same procedure was repeated with standard Ibuprofen at different concentrations of (1, 2, 4, 8 and 16 µg/ml).

The activity was measured as percentage inhibition and calculated as follows

$$\% \text{ Inhibition} = \frac{\text{Abs of control} - \text{Abs of test}}{\text{Abs of control}} \times 100 \quad (1)$$

## 4 In Vivo Pharmacological Studies

### 4.1 Preparation of Test Drug and Dose Fixation

UA was prepared by suspending in 1% CMC solution freshly before oral administration to the animals using oral feeding needle. Anti-inflammatory activity of Usnic acid was reported by Vijaykumar et al., “at doses of 25, 50 and 100 mg/kg, p.o”. [14]. Hence, the current study used 50 and 100 mg/kg, p.o of UA as low and high doses respectively.

**Table 1** Interaction Energies obtained for ligand with different molecular targets

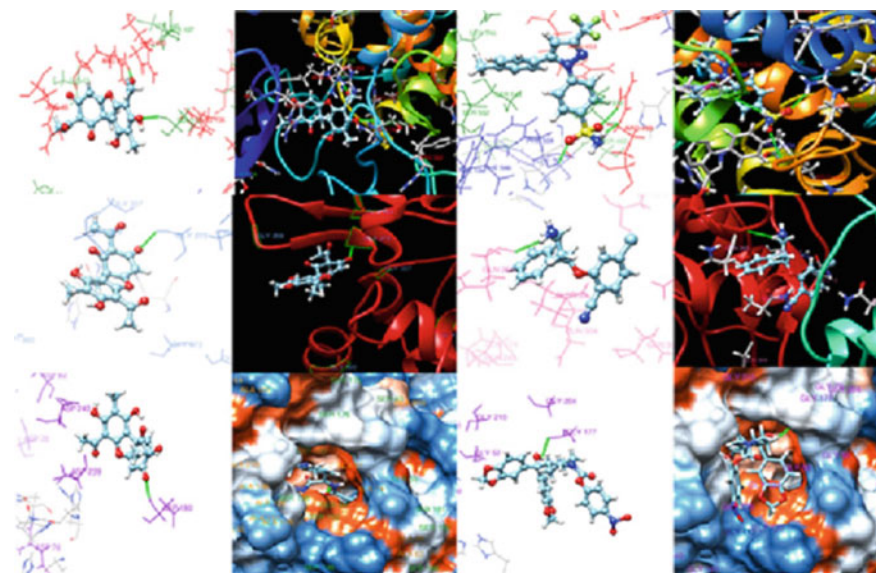
S. No.	Target (PDB ID)	Compound	Binding Energy ( $\Delta G$ ) (Kcal/mol)	Full fitness (Kcal/mol)	Interacting aminoacids	Bond distance ( $\text{\AA}$ )
1	Cyclooxygenase (COX)	Usnic acid	-7.62	-2302.99	Lig O—ARG 29 H	2.05
					Lig O—SER 112 NH	2.05
					Lig O—ARG 499 H	2.49
					Lig O—PHE 504 NH	2.48
					Lig H—SER 339 O	2.09
2	iNOS	Usnic acid	-7.28	-4112.90	Lig O—GLY 273 NH	2.57
					Lig H—GLN 265 O	2.21
					Lig O—ASP 180 NH	2.04
3	MAGL	Usnic acid	-7.52	-1122.21	Lig H—GLY 177 O	2.71
					Lig O—TYR 246 NH	2.01
4	TRPV1	Usnic acid	-7.08	-1440.11	Lig O—THR 239 H	1.49
					Lig O—TYR 246 NH	2.00
5	GABA	Usnic acid	-7.39	-1380.51	Lig O—ARG 142 H	1.98
					Lig O—ARG 142 H	2.67
					Lig O—ILE 218 NH	1.96
					Lig O—LYS 215 H	2.06
					Lig O—GLN 64 H	2.32

(continued)



**Table 1** (continued)

S. No.	Target (PDB ID)	Compound	Binding Energy ( $\Delta G$ ) (Kcal/mol)	Full fitness (Kcal/mol)	Interacting aminoacids	Bond distance ( $\text{\AA}$ )
6	Glutamate	Usnic acid	-8.38	-5044.81	Lig O—LYS 240 H	2.51
		Standard (RN8)	-9.32	-5022.39	Lig O—ALA 175 NH Lig H—GLU 13 O Lig O—SER 14 NH Lig O—TYR 199 H	2.18 2.27 2.62 2.36
7	Monoamine oxidase	Usnic acid	-7.70	-2567.43	Lig O—TRP 116 NH	2.14
					Lig O—LYS 102 H	2.05
					Lig O—LYS 102 H	2.69
		Standard (Idazoxan)	-7.37	-2516.41	Lig H—PRO 114 O Lig H—GLU 492 H Lig N—TYR 121 H	1.94 1.84 2.41



**Fig. 1** *In silico* docking studies of usnic acid and standard drug on various receptors (COX, iNOS and MAGL) as per Table 1

## 4.2 Experimental Animals

Male Swiss albino mice weighing 25–30 g were purchased from Sri Venkateswara Enterprises, Bangalore and housed under accepted environmental conditions. All the protocols used in the study were carried out in conformity with the guidelines of CPCSEA and approved by Institutional animal ethical committee with Regd. No. 1677/PO/Re/2012/CPCSEA/6, dated 3rd March 2016.

## 4.3 Evaluation of Analgesic Activity

The experimental animals were randomly divided into 4 groups ( $n = 6$ ). Group I was normal, Group II received inducer [Capsaicin ( $1.6 \mu\text{l}$ ); Glutamate ( $20 \mu\text{l}$ ); Formalin ( $20 \mu\text{l}$ )], Group III and IV were given inducer along with UA (50 and 100 mg/kg, p.o) respectively. All the experimental groups were treated as given in Table 1. For all the models, Swiss albino mice were pretreated with the test compounds or vehicle orally one hour prior to the administration of inducer.

### 4.3.1 Capsaicin Induced Analgesia

“The freshly prepared capsaicin (1.6  $\mu\text{g}$  in 20  $\mu\text{l}$  of physiological saline and ethanol (5:1, v/v) was injected into the dorsal surface of the right hind paw of mice and inducer in this model. After capsaicin injection, mice were observed individually for 15 min for their behavioral parameters like number of lickings and flinchings” as per the standard procedures [15].

### 4.3.2 Formalin-Induced Paw Licking

The formalin test was performed as previously described [16]. “The inducer of this model was 20  $\mu\text{l}$  of 5% (w/w) formalin solution, injected into the dorsal surface of right hind paw. Immediately after formalin injection, the animals were placed individually under glass beaker and observed for 15 min”. The number of lickings, bitings and flinchings were measured for a period of 0–10 min, as an indicator of nociceptive behavior.

### 4.3.3 Glutamate Induced Nociception

To evaluate glutamate induced analgesia mice were treated with test compound or standard drug orally. The test compounds/standard was administered by the oral route 1 h prior the administration of glutamate (10  $\mu\text{M}$ /paw). “Mice were injected with 20  $\mu\text{L}$  of glutamate in the sub-plantar region of the right hind paw and observed for 15 min for number of lickings, flinchings and jumpings which indicates nociceptive score” as reported earlier [17].

### 4.3.4 Tail Flick Test

After one hour of the test compounds administration (usnic acid—50 and 100 mg/kg, p.o; standard- tramadol- 10 mg/kg, p.o), the tip of tail was dipped into hot water at a depth of 5 cm maintained at  $55 \pm 1$  °C. “The response time was noted as the sudden withdrawal of the tail from the hot water and cut off time of 15 s was maintained to avoid damage to the tail. The time required for flicking of the tail was recorded to assess response to noxious stimulus in different experimental groups” as stated in earlier reports [18].

### 4.3.5 Involvement of ATP-Sensitive $\text{K}^+$ Channel Pathway

“The possible involvement of ATP-sensitive  $\text{K}^+$  channel in usnic acid mediated antinociceptive effect was evaluated” using previously described method by Mohamad et al. [19]. Fifteen minutes prior the administration of test compound,

mice were treated with glibenclamide (10 mg/kg, p.o). “Following 1 h of test compounds (usnic acid-25 and 50 mg/kg p.o) and standard drug (10 mg/kg, p.o), animals were injected with acetic acid (0.6%, i.p) and immediately placed in a polypropylene chamber. Following 5 min of acetic acid injection, the number of writhings were recorded for 30 min” as per standard protocol.

### Statistical analysis

Data was expressed as mean  $\pm$  standard deviation (SD) of 6 observations. Statistical difference was analyzed using “one-way analysis of variance (ANOVA) followed by Dunnett’s test using Graph Pad Prism version 7 software (Graph Pad Software, Inc. La Jolla, CA, USA)”. The value of  $p < 0.05$  was considered as statistically significant.

## 5 Results and Discussion

### Computational studies

Usnic acid exhibited good binding affinity towards all the selected molecular targets with interaction energies ranging from  $-7.08$  to  $-10.29$  K Cal/Mol. The results also revealed that usnic acid showed better binding affinity towards all the studied targets. The effect of usnic acid was comparable with the standard, indicating the anti-inflammatory activity reported in many animal models. Its interaction with MAGL hints that usnic acid can also act on central pain pathway targeting cannabinoid receptors. Interaction Energies obtained for ligand with different molecular targets were given in Table 1 and Fig. 1 shows the in silico docking images.

### In vitro assays of usnic acid

The enzymes released during inflammation initiate a variety of disorders. The extra cellular activity of these enzymes are proved to be related to acute or chronic inflammation. “As HRBC membrane is similar to lysosomal membrane components of the human cell, the compounds inhibiting/stabilizing the HRBC membrane were considered to possess anti-inflammatory activity. Similarly, proteinases such as trypsin and other serine proteases were reported to mediate the hydrolytic breakdown of peptide bonds in proteins” [20]. “Protein denaturation has also been well correlated with the occurrence of the inflammatory response and leads to various inflammatory diseases including arthritis” [21]. Lipoxygenases (LOXs) have been implicated in the metabolism of inflammation mediators and immune response [22]. Hence, the inhibition of these enzymes becomes imperative and therapeutic inhibition of these enzymes are approved targets and models for developing potential anti-inflammatory agents.

The inhibition of haemolysis, proteinases, protein denaturation and LOXs by UA was found to be dose dependent. Here, the effect was comparable with that of standard ibuprofen. The % inhibition of in vitro models and their  $IC_{50}$  values were given in the Table 2.

**Table 2** Effect of usnic acid on in vitro anti-inflammatory activity

S. No.	Concentration ( $\mu\text{g/ml}$ )	HRBC method		Proteinase inhibition		Albumin denaturation assay		Lipoxygenase inhibition assay	
		Ibuprofen (%inhibition)	Usonic acid (% inhibition)	Ibuprofen (%inhibition)	Usonic acid (% inhibition)	Ibuprofen (%inhibition)	Usonic acid (% inhibition)	Ibuprofen (%inhibition)	Usonic acid (% inhibition)
1.	1	15.92 $\pm$ 2.42	36.08 $\pm$ 2.34	15.05 $\pm$ 1.39	45.02 $\pm$ 2.45	33.74 $\pm$ 1.43	39.42 $\pm$ 2.34	25.85 $\pm$ 2.34	23.90 $\pm$ 2.44
2.	2	18.66 $\pm$ 3.56	43.10 $\pm$ 1.78	33.15 $\pm$ 2.12	45.09 $\pm$ 2.34	48.68 $\pm$ 2.45	44.02 $\pm$ 3.44	33.54 $\pm$ 3.23	61.90 $\pm$ 2.78
3.	4	28.45 $\pm$ 2.34	48.26 $\pm$ 3.56	69.10 $\pm$ 2.67	47.92 $\pm$ 3.56	51.87 $\pm$ 3.45	55.44 $\pm$ 2.45	39.52 $\pm$ 1.45	75.72 $\pm$ 3.55
4.	8	32.47 $\pm$ 1.56	48.50 $\pm$ 2.89	70.25 $\pm$ 3.44	58.79 $\pm$ 2.89	61.05 $\pm$ 4.56	66.32 $\pm$ 3.67	49.15 $\pm$ 3.78	86.31 $\pm$ 4.56
5.	16	45.01 $\pm$ 3.43	50.96 $\pm$ 2.88	84.20 $\pm$ 1.89	71.57 $\pm$ 3.67	70.55 $\pm$ 2.67	79.23 $\pm$ 5.45	59.37 $\pm$ 2.45	93.77 $\pm$ 5.34
	( $\mu\text{g/ml}$ )	IC <sub>50</sub> = 13.11	IC <sub>50</sub> = 17	IC <sub>50</sub> = 5.1	IC <sub>50</sub> = 4.23	IC <sub>50</sub> = 4.67	IC <sub>50</sub> = 3.54	IC <sub>50</sub> = 10.31	IC <sub>50</sub> = 1.00

All the values were expressed as mean  $\pm$  SD of triplicates

### In vivo analgesic activity

The in vivo models of analgesic activity have been selected based on the targets involved in the pain pathway. “Capsaicin induced analgesia is a relevant model since it is a selective TRPV1 agonist capable of inducing an acute nociception and neurogenic inflammation in experimental animals” [23, 24]. “Glutamate is an excitatory amino acid widely known to play major role in pain perceptions by acting through peripheral, spinal, and supraspinal sites of actions using both N-methyl-D-aspartate (NMDA) and non-NMDA receptors” [25]. The pattern of pain caused by formalin is neurogenic in nature in the first five minutes involving substance p and bradykinins. The later phase for ten minutes was mediated by peripheral nociceptive mediators [26].

Usnic acid when given orally exhibited a dose dependent antinociceptive effect on the capsaicin/glutamate/formalin induced neurogenic paw-lickings, flinching and jumpings as a response. Interestingly UA was also able to abolish biphasic response of formalin induced pain as given in Table 3. Tail flick-induced nociception was considered to be mediated by spinal mechanisms [27]. UA was also capable of influencing thermal nociception due to spinal mechanisms as it was able to enhance the threshold of pain in the tail immersion model suggesting that it acts on spinal mediated pain pathways (Table 4).

**Table 3** Effect of usnic acid in capsaicin/Glutame/formalin induced pain

Model	Parameters	Disease control	Standard (10 mg/kg, p.o)	Test low dose (50 mg/kg, p.o)	Test low dose (100 mg/kg, p.o)
Capsaicin induced model	No. of lickings	28.8 ± 3.77	9.8 ± 4.14 <sup>a</sup>	21.4 ± 4.16 <sup>a</sup>	14.6 ± 3.84 <sup>a</sup>
	No. of flinchings	17.0 ± 7.45	5.8 ± 1.30 <sup>a</sup>	15.0 ± 1.58 <sup>a</sup>	6.8 ± 1.30 <sup>a</sup>
	No. of Jumpings	13.6 ± 5.68	3.0 ± 1.34 <sup>a</sup>	11.8 ± 1.92 <sup>a</sup>	7.2 ± 2.28 <sup>a</sup>
Glutamate Induced pain	No. of lickings	30.0 ± 3.00	9.6 ± 2.07 <sup>a</sup>	19.0 ± 4.51 <sup>a</sup>	17 ± 2.70 <sup>a</sup>
	No. of flinchings	18.0 ± 4.72	3.9 ± 1.76 <sup>a</sup>	11.4 ± 2.88 <sup>a</sup>	7.6 ± 2.70 <sup>a</sup>
	No. of jumpings	17.4 ± 10.47	4.2 ± 3.19 <sup>a</sup>	15.6 ± 6.77 <sup>a</sup>	12.8 ± 5.4 <sup>a</sup>
Formalin induced pain	No. of lickings	74.2 ± 6.22	18.2 ± 2.86 <sup>a</sup>	64.2 ± 22.80 <sup>a</sup>	40.2 ± 8.29 <sup>a</sup>
	No. of flinchings	33.4 ± 4.28	1.0 ± 5 <sup>a</sup>	22.0 ± 3.54 <sup>a</sup>	20.1 ± 3.27 <sup>a</sup>
	No. of jumpings	24.6 ± 2.70	6.0 ± 3.54 <sup>a</sup>	20.8 ± 3.77 <sup>a</sup>	17.6 ± 3.65 <sup>a</sup>

Values were expressed as Mean ± SD (n = 6)

<sup>a</sup>p < 0.05, considered statistically significant when compared to the disease control

**Table 4** Effect of usnic acid in tail flick test

S. No.	Groups	Tail withdrawal latency (Time in s)
1	Disease control	4.8 ± 0.76
2	Standard (10 mg/kg, p.o)	13.2 ± 0.58 <sup>a</sup>
3	Test low dose (50 mg/kg, p.o)	11.4 ± 0.56 <sup>a</sup>
4	Test high dose (100 mg/kg, p.o)	14.8 ± 0.99 <sup>a</sup>

Values were expressed as Mean ± SD (n = 6)

<sup>a</sup>*p* < 0.05, considered statistically significant when compared to the disease control

**Table 5** Effect of usnic acid on involvement of ATP sensitive K<sup>+</sup> channel path way

S. No.	Group	Treatment	No. of writings
1	I	Acetic acid (AA) + (0.1 ml of 0.6%, i.p)	124.21 ± 9.76
2	II	Glibenclamide (10 mg/kg, p.o) + AA	116.54 ± 12.15
3	III	Ibuprofen (10 mg/kg, p.o) + AA	93.62 ± 9.64
4	IV	Ibuprofen(10 mg/kg) + Glibenclamide (10 mg/kg, p.o) + AA	92.17 ± 9.85
5	V	Usonic acid (50 mg/kg, p.o) + AA	41.67 ± 9.73 <sup>a</sup>
6	VI	Usonic acid (50 mg/kg) + Glibenclamide(10 mg/kg, p.o) + AA	99.80 ± 8.75
7	VII	Usonic acid (100 mg/kg) + AA	32.78 ± 4.06 <sup>a</sup>
8	VIII	Usonic acid (100 mg/kg) + Glibenclamide(10 mg/kg, p.o) + AA	89.20 ± 3.16

Values were expressed as Mean ± SD (n = 6)

<sup>a</sup>*p* < 0.05, considered statistically significant between groups V Vs VI and VII Vs VIII

Substantial scientific reports revealed that glibenclamide specifically blocks only the ATP-sensitive K<sup>+</sup> channels but does not affect other types like Ca<sup>2+</sup> activated and voltage-gated K<sup>+</sup> channels. Hence, measuring acetic acid induced writings in presence and absence of glibenclamide help us to identify the effect of compounds on ATP-sensitive K<sup>+</sup> channels. Interestingly, the number of writings increased in groups pre-administered with glibenclamide when compared to the groups receiving usnic acid alone. Such behavior was absent in the standard ibuprofen. The results were mentioned in the Table 5.

## 6 Conclusion

The present study evaluated antinociceptive and anti-inflammatory mechanisms of usnic acid through numerous in silico and in vitro methods whose results strongly correlate with in vivo studies. Results from in silico studies revealed the proficiency of usnic acid to bind with the peripheral and central nociceptors. The in vitro studies also

suggested the anti-inflammatory effect of usnic acid supporting the *in silico* docking reports. On other hand the *in vivo* results demonstrated that usnic acid significantly inhibited the pain thresholds mediated by both peripheral and central nociceptors. Therefore, from our results, we can conclude that amalgamation of *in silico*, *in vitro* and *in vivo* screening models will help us understand insights of therapeutic activity, establishing the safety and efficacy of interventional drugs.

## References

1. G. Dothel, V. Vasina, G. Barbara, F. De Ponti, Animal models of chemically induced intestinal inflammation: predictivity and ethical issues. *Pharmacol. Ther.* **139**, 71–86 (2013)
2. J.C. Madden, M. Hewitt, K. Przybylak, R.J. Vandebriel, A.H. Piersma, M.T. Cronin, Strategies for the optimisation of *in vivo* experiments in accordance with the 3Rs philosophy. *Regul. Toxicol. Pharmacol.* **63**, 140–154 (2012)
3. K.L. Chapman, H. Holzgrefe, L.E. Black, M. Brown, G. Chellman, C. Copeman, J. Couch, S. Creton, S. Gehen, A. Hoberman, L.B. Kinter, Pharmaceutical toxicology: designing studies to reduce animal use, while maximizing human translation. *Regul. Toxicol. Pharmacol.* **66**(1), 88–103 (2013)
4. M. Pellegatti, Preclinical *in vivo* ADME studies in drug development: a critical review. *Expert Opin. Drug Metab. Toxicol.* **8**, 161–172 (2012)
5. K.H. Bleicher, H.J. Böhm, K. Müller, A.I. Alanine, A guide to drug discovery: hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2**, 369–372 (2003)
6. C.M. Masimirembwa, R. Thompson, T.B. Andersson, *In vitro* high throughput screening of compounds for favorable metabolic properties in drug discovery. *Comb. Chem. High Throughput Screening* **4**, 245–263 (2001)
7. R. Deraedt, S. Jouquey, F. Delevallee, M. Flahaut, Release of prostaglandins E and F in an algogenic reaction and its inhibition. *Eur. J. Pharmacol.* **61**, 17–24 (1980)
8. P.K. Mukherjee, P.J. Houghton (eds.), *Evaluation of Herbal Medicinal Products: Perspectives on Quality, Safety and Efficacy* (Pharmaceutical press, London, 2009), pp. 399–401
9. A.A. Araújo, M.G. De Melo, T.K. Rabelo, P.S. Nunes, S.L. Santos, M.R. Serafini, M.R. Santos, L.J. Quintans-Júnior, D.P. Gelain, Review of the biological properties and toxicity of usnic acid. *Nat. Prod. Res.* **29**(23), 2167–2180 (2015)
10. U.A. Shinde, A.S. Phadke, A.M. Nair, A.A. Mungantiwar, V.J. Dikshit, M.N. Saraf, Membrane stabilizing activity—a possible mechanism of action for the anti-inflammatory activity of *Cedrus deodara* wood oil. *Fitoterapia* **70**, 251–257 (1999)
11. J. Sadique, W.A. Al-Rqobahs, E.I. Bughaith, A.R. Gindi, The bioactivity of certain medicinal plants on the stabilization of RBC membrane system. *Fitoterapia* **60**, 525–532 (1989)
12. O.O. Oyedapo, A.J. Famurewa, Antiprotease and membrane stabilizing activities of extracts of *Fagara zanthoxyloides*, *Olax subscorpioides* and *Tetrapleura tetraptera*. *Int. J. Pharmacognosy* **33**, 65–69 (1995)
13. S. Sakat, A.R. Juvekar, M.N. Gambhire, *In vitro* antioxidant and anti-inflammatory activity of methanol extract of *Oxalis corniculata* Linn. *Int. J. Pharm. Pharm. Sci.* **2**, 146–155 (2010)
14. C.S. Vijayakumar, S. Viswanathan, M.K. Reddy, S. Parvathavarthini, A.B. Kundu, E. Sukumar, Anti-inflammatory activity of (+)-usnic acid. *Fitoterapia* **71**, 564–566 (2000)
15. J. Sawynok, A. Reid, J. Meisner, Pain behaviors produced by capsaicin: influence of inflammatory mediators and nerve injury. *J. Pain* **7**, 134–141 (2006)
16. S. Hunskaar, K. Hole, The formalin test in mice: dissociation between inflammatory and non-inflammatory pain. *Pain* **30**, 103–114 (1987)
17. F.C. Meotti, I. dos Santos Coelho, A.R.S. Santos, The nociception induced by glutamate in mice is potentiated by protons released into the solution. *J. Pain.* **11**, 570–578 (2010)



18. A.W. Bannon, A.B. Malmberg, Models of nociception: hot-plate, tail-flick, and formalin tests in rodents. *Curr. Protoc. Neurosci.* **41**, 8–9 (2007)
19. A.S. Mohamad, M.N. Akhtar, S.I. Khalivulla, E.K. Perimal, M.H. Khalid, H.M. Ong, S. Zareen, A. Akira, D.A. Israf, N. Lajis, M.R. Sulaiman, Possible participation of nitric oxide/cyclic guanosine monophosphate/protein kinase C/ATP-Sensitive K<sup>+</sup> channels pathway in the systemic antinociception of flavokawin B. *Bas. Clin. Pharmacol. Toxicol.* **108**(6), 400–405 (2011)
20. G.A. Joanitti, S.M. Freitas, L.P. Silva, Proteinaceous protease inhibitors: structural features and multiple functional faces. *Curr. Enzym. Inhib.* **2**, 199–217 (2006)
21. Y. Mizushima, M. Kobayashi, Interaction of anti-inflammatory drugs with serum proteins, especially with some biologically active proteins. *J. Pharm. Pharmacol.* **20**, 169–173 (1968)
22. C. Pergola, O. Werz, 5-Lipoxygenase inhibitors: a review of recent developments and patents. *Expert Opin. Ther. Pat.* **20**, 355–375 (2010)
23. M.S. Gold, G.F. Gebhart, Nociceptor sensitization in pain pathogenesis. *Nat. Med.* **16**, 1248–1250 (2010)
24. N.E. Saade, C.A. Massaad, C.I. Ochoa-Chaar, S.J. Jabbur, B. Safieh-Garabedian, S.F. Atweh, Upregulation of proinflammatory cytokines and nerve growth factor by intraplantar injection of capsaicin in rats. *J. Physiol.* **545**, 241–253 (2002)
25. A. Beirith, A.R. Santos, J.B. Calixto, Mechanisms underlying the nociception and paw oedema caused by injection of glutamate into the mouse paw. *Brain Res.* **924**, 219–228 (2002)
26. A.R.S. Santos, J.B. Calixto, Further evidence for the involvement of tachykinin receptor subtypes in formalin and capsaicin models of pain in mice. *Neuropeptides* **31**, 381–389 (1997)
27. T.S. Jensen, D.F. Smith, Dopaminergic effects on tail-flick response in spinal rats. *Eur. J. Pharmacol.* **79**, 129–133 (1982)

# Herbal Tea Treatment of Oligomenorrhea Condition with Hibiscus Rosa-Sinensis and Carica Papaya



G. Sreesha and D. Sai Prasanna

**Abstract** Oligomenorrhea is a gynaecological medicinal situation of irregularities through menstruation that affects with daily activities and others. Numerous things can cause asymmetrical periods. Changes in the hormones estrogens and progesterone can disturb the regular pattern of period. That's why young adolescents going through puberty commonly have irregular periods. The occurrence of Oligomenorrhea between teenage females ranges from 60 to 73%. Many adolescent girls report limitations on daily activities, such as sports events. Based on the symptoms of Oligomenorrhea the subjects were purposely selected. The aim of the research was to reduce this problem by supplementing them with unripe papaya and hibiscus flowers. A total number of 30 subjects were purposively selected for experimentation. The information regarding the Socio-economic status, food habits, anthropometric measurements and clinical information etc. were collected by using the questionnaire. The research consists of three experimental groups, which are group 1 experimental, group 2 experimental and group 3 control. Each experimental group consists of 10 members of subjects. Experimental group-1 receives hibiscus tea, experimental group-2 receives unrefined papaya tea. The supplementation was continued two months, experimental group-1 showed changes in the clinical symptoms after hibiscus tea supplementation.

**Keywords** Hibiscus · Menstrual pain · Oligomenorrhea · Papaya · Tea supplementation

---

G. Sreesha (✉) · D. Sai Prasanna  
Department of Home Science, Sri Padmavati Mahila Visvavidyalayam,  
Tirupati, India  
e-mail: [sireeshaguttapalam@gmail.com](mailto:sireeshaguttapalam@gmail.com)

D. Sai Prasanna  
e-mail: [saiprasanna739@gmail.com](mailto:saiprasanna739@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_23](https://doi.org/10.1007/978-3-030-46939-9_23)

## 1 Introduction

Oligomenorrhea is a condition where have rare menstrual periods. It happens in ladies of childbearing age. Some variety in monthly cycle is ordinary, yet a lady who routinely goes over 35 days without discharging might be determined to have Oligomenorrhea [1]. Oligomenorrhea with a commonness of 12–15.3% in various investigations around the globe is one of the most well-known kinds of menstrual draining issue. In ongoing decades, because of changes in way of life, weight, low physical action, unfortunate nourishment, and passionate pressure, the commonness of amenorrhea and Oligomenorrhea has expanded extensively. Among a few etiologic elements, polycystic ovarian infection (PCOD) is the most significant fundamental factor for Oligomenorrhea [2]. Ladies with Oligomenorrhea may have the accompanying manifestations, menstrual periods at interims of over 35 days, abnormally light menstrual stream, sporadic menstrual periods with eccentric stream, trouble considering.

Welt et al. [3] in his release, Oligomenorrhea is a medicinal term to portray changes in recurrence of menstrual bleeds where already periods were unsurprising. Indications of Oligomenorrhea are the place drains happen at more prominent interims than 35 days with just 4 to 9 periods for each year. A few ladies with in the investigation of unpredictable periods with identified with corpulence has demonstrated that stoutness at last lead to an expansion in ripeness issues and sporadic periods. Pregnant hefty ladies may cause sporadic periods in their little girls. It is to the greatest advantage of hopeful moms or ladies who plan on having kids later on to keep up a sound weight. This is only one deterrent strategy to diminish the opportunity of unpredictable periods. Oligomenorrhea can likewise be incited by enthusiastic and physical pressure brought about by a dietary issue like anorexia nervosa. Said and Mettwaly [4] have directed an examination to improve way of life among nursing understudies with respect to menstrual issue through an instructive preparing program.

**Hibiscus Rosa-sinensis** (hibiscus flowers) are used for several medicinal uses in different countries. Hibiscus flowers are astringent, demulcent, emollients, refrigerants, constipating, hypoglycaemic, aphrodisiac, and used for treating alopecia, burning sensation in the body, diabetes and menstrual disorders [5].

Hibiscus flower regulates the oestrogen and progesterone balance within the body, thus helping the monthly cycle to be regular and balanced. These flowers have an anti-oestrogenic quality, which helps to regularise the oestrogen/progesterone balance and initiate the menstrual flow in time. Hibiscus flower are very effective and wonder to perform against Oligomenorrhea [6].

Wong [7] in his overview both creature and human models have exhibited that concentrates or implantations influence atherosclerosis systems, glucose, lipids, circulatory strain, diabetes, metabolic disorder and twofold visually impaired control. Dubois [8] the blooms and foundations of the hibiscus plant can likewise influence estrogen levels, and ladies ought to be wary when utilizing them.

Unripe papaya may likewise be taken each day to fix a wide range of menstrual inconsistencies. papaya is most appropriate for sporadic periods identified with menopause [9]. Adaikan et al. [10] results recommend that typical utilization of ready papaya during pregnancy may not represent any huge threat. The unripe or semi-ready papaya (which contains high grouping of the latex that produces checked uterine constrictions) could be risky in pregnancy.

Aravind et al. [9] Carica papaya smooth juice is separated, dried and utilized as a biting gum for stomach related issues, toothpaste and meat tenderizers. It additionally contains numerous organic dynamic mixes including chymopapain and papain which is the fixing that guides stomach related framework, and again utilized in treatment of joint inflammation. Meera and Ugendra [11] They affirm that uterine stimulant movement of watery concentrate of unripe carica papaya natural product. Studies show that hibiscus flowers and unripe papaya are widely used to treat the different health problems. So papaya and hibiscus supplementation for adolescent girls were selected to control Oligomenorrhea and associated risk factors. Hibiscus flowers and unripe papaya are helpful in hormone balances. Due to presence of an anti-estrogenic quality, which helps to regularize the estrogen/progesterone balance and initiate the menstrual flow in time and helps contract muscle fibers in the uterus that induce periods.

## 2 Materials and Methods

The subjects from Sri Padmavati Mahila Visvavidyalayam who are from various parts of Andhra Pradesh and they fall under the age range of 20–23 years. Based on the symptoms of Oligomenorrhea the subjects were purposely selected. A total number of 30 subjects were selected for supplementation programme.

### **Selection of tools and techniques**

Based on the questionnaire general information, clinical information and dietary information were collected. All the selected subjects were asked to fill a questionnaire which contain details about their name, age, height, weight, education, occupation, family, income, details about their symptoms and their respective time of occurrence, duration back pain, joint pains, light bleeding menstrual flow, hot flashes, heavy white vaginal discharge, vaginal dryness and other symptoms etc. dietary habits, likes and dislikes were collected by interviewing them.

### **Anthropometry method**

Anthropometric measurement like height, weight, hip and waist recorded to estimate the nutritional status of selected samples. Body mass index (BMI) is a person's weight in kilograms divided by his or her height in meters square.

$$\text{BMI} = \text{actual weight in kg/height in m}^2.$$

Waist-to-waist ratio: WHR is used as a measure of obesity, which in turn is a possible indicator of other more serious health conditions. WHO 2008 states that



**Fig. 1** Hibiscus instant tea bags

abdominal obesity is defined as a waist-hip greater than 1.00 for men and greater than 0.85 for women or a body mass index greater than 30.

### **Hibiscus instant Tea bags**

- a. Collected the hibiscus flowers and washed them clean, and dry.
- b. Soak flowers in lemon juice for 5–10 min and dry them in shade (shadow drying) for 1 week.
- c. After drying make them into coarse powder as tea powder.
- d. Dry hibiscus powder was packed in empty hygiene tea bags in required amounts (0.55 gm).
- e. Keep them in a mug, add 200 ml boiling water, steep it for 2–4 min, add sugar or honey if desired.
- f. Hibiscus tea has an intense pink colour due to the petals used to prepare the drink. Beer is often enjoyed as a hot herbal tea or as a refreshing iced tea.
- g. Hibiscus powder is suitable for consumption without the need for further processing. Hence it was selected for supplementation. Figure 1 shows the hibiscus tea.

### **Unripe papaya instant tea**

- a. First, collect the unripe papaya and wash them clean.
- b. after washing, the skin of papaya is peeled and seeds were removed.
- c. make the fruit into fine pieces and dry them under sun (sun drying) for 2–3 weeks.
- d. after drying make them into coarse powder as tea powder.
- e. Dry powder was packed in empty hygiene tea bags in required amounts (0.55 g).
- f. Put them in a tea bag, bring out favourite mug, add 200 ml of boiling water, steep it for 2–4 min, add honey if desired. Hence it was selected for supplementation. Figure 2 shows the unripe papaya tea.

### **Supplementation of developed products to subjects**

Thirty members divided into three experimental groups. Each experimental group consists of 10 members of subjects. First group act as control group, it consists of 10 subjects, 5 of them with normal body weight and Oligomenorrhoea condition and remaining 5 of them with obese and Oligomenorrhoea condition. Second group acts as experimental group-1, it consists of 10 subjects, with obese and Oligomenorrhoea condition. Third group acts as experimental group-2, it consists of 10 subjects with



**Fig. 2** Papaya tea instant tea bags

normal body weight and Oligomenorrhea condition. The supplementation period was two months. The tea powder dosage was 0.55 gm/day in the morning time subjects were consumed.

### 3 Results and Discussion

**Age:** The data in Table 1 shows the distribution of the subjects according to age. A majority of the Oligomenorrhea subjects were in the age of 21–22 years. Remaining Oligomenorrhea subjects were in the age group of 20–21 and 20–23 years respectively. It is observed that in both experimental groups and control groups majority of them are suffering with Oligomenorrhea, this may be due to stress, irregular food habits and high education prevalence of Oligomenorrhea is more in age group of 21–22 years. These results are on par with the results of **Ogden**, [12] who reported that hormonal imbalances and stress are the factors for high prevalence of Oligomenorrhea in young ladies in the age group of 20–25 years than other age groups. The incidence of Oligomenorrhea in women of reproductive age is approximately 50–60%. It is greatest in women in their late teens to early 20s.

**Weight:** The data in Table 2 depicts that sample distribution according to weight in kgs. In the exp group-1 20% of the subjects showed change in their weights.

**Height:** Through the majority of Oligomenorrhea subjects is normal weight range but due to their, irregular food habits and physical inactivity they have developed Oligomenorrhea. The present study results are on par with the results of [13] reported

**Table 1** Percentage distribution of experimental groups and control group according to age

Age (years)	Experimental group-1 (n = 10)	Experimental group-2 (n = 10)	Control group (n = 10)
20–21	1(10)	3(30)	–
21–22	6(60)	4(40)	7(70)
22–23	3(30)	3(30)	3(30)

Figure in () indicates percentage

**Table 2** Percentage distribution of experimental and control subjects of Oligomenorrhoea according to weight before and after supplementation

Weight in kg	Experimental group-1 (n = 10)		Experimental group-2 (n = 10)		Control group (n = 10)	
	Before	After	Before	After	Before	After
40–50	2(20)	2(20)	6(60)	6(60)	–	–
50–60	2(20)	4(40)	4(40)	4(40)	5(50)	5(50)
60–70	4(40)	2(20)	–	–	3(30)	3(30)
70–80	2(20)	1(10)	–	–	2(20)	2(20)

Figure in () indicates percentage

that improper food habits lead to many health disorders which includes Oligomenorrhoea. The data in Table 3 indicates that percentage distribution of sample according to height in cm. In both experimental groups and control group majority (60–70%) of Oligomenorrhoea subjects are under the height range of 145–155 cm.

**BMI (Body Mass Index):** The data in Table 4 illustrates that percentage distribution of sample according to BMI, it reveals that in experiment group-1 of subjects supplementing with hibiscus tea, 60% were overweight before supplementation but after supplementation reduced to 50%. Remaining 20% were with normal BMI before and after supplementation. Other 20% were with low BMI before supplementation but it increased after supplementation to 30%. Majority of both the experimental and

**Table 3** Percentage distribution of experimental groups and control group according to height

Height in cm	Experimental group-1 (n = 10)	Experimental group-2 (n = 10)	Control group (n = 10)
145–155	6(60)	9(90)	7(70)
156–175	4(40)	1(10)	3(30)

Figure in () indicates percentage

**Table 4** Percentage distribution of experimental and control groups according to Body Mass Index before and after supplementation

BMI grades classification (kg/m <sup>2</sup> )	No. of samples					
	Experimental group-1 (n = 10)		Experimental group-2 (n = 10)		Control group (n = 10)	
	Before	After	Before	After	Before	After
Underweight ( $\leq 18.5$ )	2(20)	3(30)	4(40)	4(40)	5(50)	5(50)
Normal weight (18.5–24.9)	2(20)	2(20)	4(40)	4(40)	4(40)	4(40)
Over weight and above ( $\geq 25$ )	6(60)	5(50)	2(20)	2(20)	1(10)	1(10)

Figure in () indicates percentage

control groups subjects were having normal BMI, due to their improper food habits they might have develop menstrual disorders such as Oligomenorrhea.

The present study results are on par with the results of [14] who stated that due to faulty food habits and high junk food consumption develops menstrual disorders and Oligomenorrhea is prevailing among adolescents.

WHR is used as a measurement of obesity, which in turn is a possible indicator of other more serious health conditions. The WHO states that abdominal obesity is defined as waist-hip above 1.00 for males and above 0.85 for females or a body mass index above 30 [15]. The data in Table 5 illustrates that percentage distribution of sample according to waist and hip ratio, it reveals that in experiment group-1 supplementation with hibiscus tea subjects 90% were with obesity before supplementation and it reduced to 40% after supplementation. In experiment group-2, 20% of the overweight subjects shifted into normal category. There was no changes observed in control group. According to WHO, waist and hip ratio in women, acts as indicator for health conditions especially in menstrual cycle disorders, fertility and Oligomenorrhea conditions.

**Clinical symptoms:** Table 6 shows the clinical symptoms of experimental and control groups before and after supplementation. Both the experimental and control group subjects were experiencing the symptoms like light bleeding, hot flashes, white bleeding and vaginal dryness. Majority in subjects of Oligomenorrhea consist of 100% of white bleeding in every subject, it is very common symptom in adolescents. Due to low protein intake at the maturation age leads to irregular periods. The symptoms were constant in control group. After supplementation with hibiscus tea subjects showed changes in their symptoms (severe to mild). 50% subjects were reduced the symptoms like light bleeding and hot flashes, also 40% subjects were reduces the symptoms like white bleeding and vaginal dryness.

**Menstrual cycle length:** The data in Table 7 describes about the menstrual cycle length of the experimental groups, before supplementation 50% of the experimental group-1 subjects fall under the >35 days menstrual cycle length, after supplementation the percentage was decreased to 20%. It means regularizes the period menstrual cycle length days correctly. Due to presence of anti estrogenic property in hibiscus it balances the hormone levels and regularizes the cycle lengths.

**Table 5** Percentage distribution of experimental and control groups according to waist and hip ratio before and after supplementation

Waist and hip ratio valves	No. of samples					
	Experimental group-1 (n = 10)		Experimental group-2 (n = 10)		Control group (n = 10)	
	Before	After	Before	After	Before	After
Normal weight (<0.80 cm)	–	1(10)	4(40)	6(60)	4(40)	4(40)
Over weight(0.80–0.84 cm)	1(10)	5(50)	4(40)	2(20)	3(30)	3(30)
Obesity (>0.85 cm)	9(90)	4(40)	2(20)	2(20)	3(30)	3(30)

Figure in () indicates percentage



**Table 6** Clinical symptoms of experimental and control groups before and after supplementation

Clinical symptoms	No. of samples					
	Experimental group-1 (n = 10)		Experimental group-2 (n = 10)		Control group (n = 10)	
	Yes	No	Yes	No	Yes	No
Light bleeding before	8(80)	2(20)	7 (70)	3(30)	7(70)	3(30)
Light bleeding after	5(50)	5(50)	3(30)	7(70)	7(70)	3(30)
Hot flashes before	7(70)	3(30)	2(20)	8(80)	4(40)	6(60)
Hot flashes after	1(10)	9 (90)	2(20)	8(80)	4(40)	6(60)
White bleeding before	10(100)	–	10(100)	–	9(90)	1(10)
White bleeding after	6(60)	4(40)	10(100)	–	9(90)	1(10)
Vaginal dryness before	9(90)	1(10)	5(10)	1(50)	4(40)	6(60)
Vaginal dryness after	4(40)	6(60)	5(10)	1(50)	4(40)	6(60)

Figure in () indicates percentage

**Table 7** Percentage distribution of sample according to menstrual cycle length (days) before and after supplementation

Menstrual cycle length	No. of samples					
	Experimental group-1 (n = 10)		Experimental group-2 (n = 10)		Control group (n = 10)	
	Before	After	Before	After	Before	After
≤20	2(20)	1(10)	2(20)	2(20)	4(40)	4(40)
21–35	3(30)	7(70)	5(50)	4(40)	3(30)	3(30)
>35	5(50)	2(20)	3(30)	4(40)	3(30)	3(30)

Figure in () indicates percentage

No changes were observed in experimental-2 and control group when comparing to before and after supplementation. Any variation from this, i.e. <24 days or >35 days is considered as irregular [16].

**Menstrual flow rate (days):** The data in Table 8 describes that in experimental group-1 supplementation with hibiscus tea, 90% of Oligomenorrhea subjects were suffering menstrual flow lesser than 2 days before supplementation and after supplementation 60% subjects menstrual flow was increased to 3–5 days. On an average, a period lasts between 3 and 5 days when in regular condition. If periods are irregular, women's were suffering menstrual flow lesser than 2 days and sometimes between 5 and 7 days or greater than 7 days causes Oligomenorrhea.

**Dietary food habits:** The data in Table 9 reveals the dietary habits of the selected subjects. In both the exp and control groups majority of the subjects are consuming non vegetarian foods. Due to high intake of non vegetarian foods obesity may occur, which in turn leads to irregular period problems.

**Table 8** Percentage distribution of experimental groups and control group according to menstrual flow (days) before and after supplementation

Menstrual flow	No. of samples					
	Experimental group-1 (n = 10)		Experimental group-2 (n = 10)		Control group (n = 10)	
	Before	After	Before	After	Before	After
≤2	9(90)	4(40)	8(80)	8(80)	6(60)	6(60)
3-5	-	6(60)	2(20)	2(20)	4(40)	4(40)
5- ≥ 7	1(10)	-	-	-	-	-

Figure in () indicates percentage

**Table 9** Dietary habits of the selected university girls

Dietary habits	Experimental group-1 (n = 10)	Experimental group-2 (n = 10)	Control group (n = 10)
Vegetarian	2(20)	-	4(40)
Non vegetarian	8(80)	10(100)	6(60)

Figure in () indicates percentage

Table 10 shows the mean, standard deviation and paired t-test values of the anthropometric measurements of the Oligomenorrhea exp and control groups after supplementation. In experimental group-1 showed significant weight changes ( $p < 0.01$ ). For waist and hip ratio values, it showed significant changes ( $p < 0.05$ ). Significant changes were not observed in BMI values. In experimental group-2, for weight it showed significant changes ( $p < 0.05$ ). No significant changes were observed in waist hip ratio and BMI values. No significant changes were observed in control group, for weight, waist hip ratio and BMI values. The results after supplementation of hibiscus tea, daily early in mornings for 2 months, showed significant results that regulates the periods in experimental group 1 with completely minimizing the signs of symptoms.

The results after supplementation of papaya tea, daily early in mornings for 2 months, showed minimum results that regulates the periods in experimental group 2 that just reduces the signs of symptoms like white bleeding and light bleeding. In control group, a significant increase in irregular periods that mild to severe.

## 4 Summary and Conclusion

Hence it can be totally concluded that hibiscus flower tea totally regularize the periods and minimize the signs of symptoms and balances the hormone levels (estrogens and progesterone) due to presence of anti-estrogenic property to it. So it can be consumed any from to regularize the periods. It shows significant changes ( $p < 0.01$ ). Unripe

**Table 10** Mean, standard deviation and paired t test of anthropometric measurements changes of the Oligomenorrhea subjects before and after supplementation

Parameters	Experimental group-1 (n = 10)			Experimental group-2 (n = 10)			Control group (n = 10)			t-values
	Before	After	t-values	Before	After	t-values	Before	After	t-values	
Weight (kg)	60.4 ±11.0	57.5 ±10.9	6.201**	48.88 ±6.76	48.56 ±6.781	2.272*	51.4 ±5.6	51.4 ±5.6214	1.000NS	
Waist and hip ratio	0.86 ±0.06	0.844 ±0.0554	2.713*	0.826 ±0.038	0.824 ±0.036	0.963NS	0.859 ±0.05	0.859 ±0.0533	0.135NS	
BMI	26.06 ±4.42	24.455 ±4.336	0.795NS	22.28 ±4.13	21.97 ±4.053	2.042NS	23.49 ±5.09	23.49 ±5.098	1.44NS	

\*\* P < 0.001 by paired t-test

\*P < 0.05 by paired t-test

papaya fruit has fibroids helps contract muscle fibers in the uterus that induce periods when it taken only in the form of fruit. So papaya tea wouldn't show any significant results on the Oligomenorrhea subjects.

## References

1. John, Shelia, *The Art of Natural Family Planning*, 2nd edn. (1996), p. 92
2. M. Yavari, S. Rouholamin, M. Tansaz, S. Esmaeili, Herbal treatment of oligomenorrhea with *Sesamum indicum* L.: a randomized controlled trial. *Galen Med. J.* **5**(3), 114–121 (2016)
3. C. Welt et al., Serum inhibin B in polycystic ovary syndrome: regulation by insulin and luteinizing hormone. *J. Clin. Endocrinol. Metab.* **87**(12), 5559–5565 (2002)
4. A.R. Said, M.G. Mettwaly, Improving life style among nursing students regarding menstrual disorders through an educational training program. *Int. J. Nurs. Sci.* **7**, 35–43 (2017). <https://doi.org/10.5923/j.nursing.20170702.01>
5. K. Puro, R. Sunjukta, S. Samir, S. Ghatak, I. Shakuntala, A. Sen, Medicinal uses of Roselle plant (*Hibiscus sabdariffa* L.): a mini review. *Indian J. Hill Farming* **27**(1), 81–90 (2014)
6. <https://afternoontoreads.com/benefits-risks-hibiscus-tea/>
7. C. Wong, *Health Benefits of Hibiscus* (2019). <https://www.verywellhealth.com/health-benefits-of-hibiscus-tea-89620>
8. S. Dubois, *Hibiscus Tea and Estrogen* (2017). <https://www.livestrong.com/article/544564-hibiscus-tea-and-estrogen/>
9. G. Aravind, D. Bhowmik, S. Duraivel, G. Harish, Traditional and medicinal uses of *Carica papaya*. *J. Med. Plants Stud.* **1**(1), 7–15 (2013)
10. G. Adaikan et al., Papaya consumption is unsafe in pregnancy: fact or fable? Scientific evaluation of a common belief in some parts of Asia using a rat model. *Br. J. Nutr.* **88**(2), 199–203 (2002)
11. Meera, Ugendra, Effect of carica papaya on uterus. *Int. J. Res. Ayurveda Pharm.* **4**(3) (2013)
12. P. Ogden, K. Minton, C. Pain, *Norton Series on Interpersonal Neurobiology. Trauma and the Body: A Sensorimotor Approach to Psychotherapy* (W W Norton & Co, New York, NY, US, 2006)
13. G.U. Liepa, A. Sengupta, D. Karsies, Polycystic ovary syndrome (PCOS) and other androgen excess-related conditions: can changes in dietary intake make a difference? *Nutr. Clin. Pract.* **23**(1), 63–71 (2008)
14. M. Barnard, Weight and body fat distribution in adolescent girls arch discoid. *Arch. Dis. Child.* **77**, 381–383 (2000)
15. C.H. Cheng, C.C. Ho, C.F. Yang, Y.C. Huang, C.H. Lai, Y.P. Liaw, Waist-to-hip ratio is a better anthropometric index than body mass index for predicting the risk of type 2 diabetes in Taiwanese population. *Nutr. Res.* **30**, 585–593 (2010)
16. J.A. Boyle, H.J. Teede, Irregular menstrual cycles in a young woman. *CMAJ.* **186**(11), 850–852 (2014)

# Distribution and Evidential Incidence of Oral Microflora Among Dental Caries Infected 3–19 Year Old in Allahabad, India—A Pilot Study



T. Jesse Joel, S. Sandeep Singh, and P. W. Ramteke

**Abstract** The routine casual life of spendthrifts in established countries and their unthinkable wayward life not much enjoyed by developing or underdeveloped countries throw a reality-check of the disease under study. Oral health influences the general quality of life and poor oral health is linked to chronic conditions and systemic diseases. Approximately 80% of the world inhabitants rely on traditional medicine for their primary health care and plants also play an important role in the health care system of the remaining 20% of the population. Here our major objective is to correlate this multifactorial disease with simple “non-risky” habits of any population unaware of the impact caused by dental caries. Teeth related problems are ascertained to be about 50–60% and very few Indian studies have been carried out. In the present pilot study, out of the total 202 patients, the distribution of the Oral flora was 94.18% among Gram-positive and 5.82% among Gram-negative microorganisms respectively.

**Keywords** Oral health · *Streptococcus mutans* · Dietary habits · Distribution · Evidence · Oral microflora · Gram positive · Gram negative

---

T. Jesse Joel (✉) · S. Sandeep Singh  
Department of Biotechnology, Karunya Institute of Technology and Sciences,  
Coimbatore, Tamil Nadu, India  
e-mail: [jessejoel@karunya.edu](mailto:jessejoel@karunya.edu)

S. Sandeep Singh  
e-mail: [drsandeep13@gmail.com](mailto:drsandeep13@gmail.com)

T. Jesse Joel  
Sahaj Dental Clinic, Allahabad, Uttar Pradesh, India

P. W. Ramteke  
Department of Biological Sciences, Sam Higginbottom Institute of Agriculture  
Technology and Sciences, Allahabad, India  
e-mail: [pwranteke@gmail.com](mailto:pwranteke@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_24](https://doi.org/10.1007/978-3-030-46939-9_24)

## 1 Introduction

The metabolism of *Streptococcus mutans* actually improves, as the proton motive system used to transport nutrients through its cell wall in environments of low pH or high glucose concentration is modulated by hydrogen ion content, which increases with acidity [1]. The microbes thrive in their respective niches and bring about a perfect balance within their habitat i.e. mouth [2].

Early Childhood Caries (ECC) or Nursing bottle Caries is a persistent entity and it is because of the flamboyant lifestyle that individuals indulge in the present day [3] where a child is not admonished when he pesters to be given a sugar based eat or even a drink. Though it is a fact that there are 700 odd microbes growing in one's mouth, it is alarming to note that more than half only have been identified [4]. Many species in the viridans group to which *Streptococcus mutans*. Belong hemolyze blood and come under the "Viridae" family.

Years of global work relating to this behavior has ascertained its etiology in less than 300 cases and the plaque, which is a consequence of biofilm production is obtained from the occlusal pits and fissures, and lower left first molars. *Streptococcus mutans* and total streptococci were counted using mitis-salivarius agar. The pits and fissures were the major source of *Streptococcus mutans* [5, 6]. Serological examinations of *Streptococcus mutans* revealed the existence of several serovars, now designated as a-h [7–9]. These are opportunistic pathogens because they cause infections when the time is right and the environment is favorable. *Streptococcus mutans*, a member of viridans group of Sherman [10] and the most virulent of these species has been found to be the initiator of most dental caries, and which is a transmissible bacterium that can be transmitted both horizontally and vertically [11–13] has been definitely established due to its pathogenic cause of cavities, as a major etiological agent of dental caries.

The World Health Organization (WHO) has identified 'Twelve' as the major risk age-group, and therefore the present work is steered in this direction. The major objective was to correlate this multifactorial disease with simple "non-risky" habits of any population unaware of the impact caused by dental caries [14].

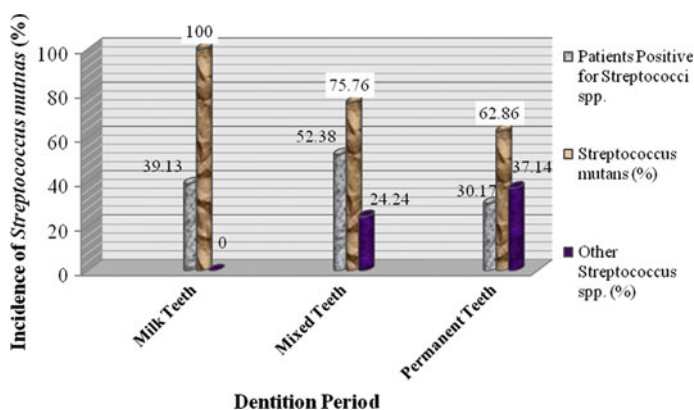
## 2 Materials and Methods

### 2.1 Materials

The various laboratory glassware and plasticware used were procured from Borosil Glass Works Ltd. (Mumbai), Merck India Head Office (Mumbai) and Polylab (A.K. Scientific Industries, Delhi) respectively. The working analytical grade chemicals were obtained from HiMedia (Mumbai), Merck (Mumbai), Sigma-Aldrich (USA), Bangalore Gene and Loba Chemie Pvt. Ltd. (Mumbai).

**Table 1** Incidence of *Streptococcus* spp. in patients with dental caries

No. of patients	Patients positive for <i>Streptococcus</i> spp. (%)	Distribution of <i>Streptococcus</i> spp.	
		<i>Streptococcus mutans</i>	Other <i>Streptococcus</i> spp.
114	<b>66(57.89)</b>	<b>53(80.30)</b>	<b>13(19.69)</b>

**Fig. 1** Incidence of *Streptococcus mutans* with respect to dentition period

## 2.2 Study Population

This Pilot study included 202 patients who came to Sahaj Dental Clinic with dental caries and among them, children aged 3–19 years ( $n = 114$ ) who were attending dental camp at Ethel Higginbottom School, Allahabad, Uttar Pradesh. Children (aged 1–12) were Eighty six (86) and an adolescent group (aged 13–15) were Twenty eight (28) as represented in Table 1. This Pilot survey study was carried out between June and December. The children and teens (Aged 3–19) were screened initially for dentition status and categorized into Primary, mixed and permanent dentition state (Fig. 1).

## 2.3 Sample Procurement

The swab sample was chosen as the most suitable specimen for the isolation of the cariogenic agent [15]. All the equipment used by the dental expert/surgeon was properly surface sterilized using disinfectants like Savlon in the ratio of 2:8 dilutions. The children were made to sit comfortably on a class stool and in a room well-lit by daylight.

The teachers were requested to help the children rinse their mouths completely before coming for inspection. If immediate treatment was required the children were

asked to meet the dentist the same evening in the Clinic. The free dental check-up was not complete until the report was submitted to the school Principal. Microbial samples were obtained from 114 children/patients (Table 1) who attended a free dental camp arranged by the University affiliated Dental Clinic in Ethel Higginbottom School, Allahabad, UP. Prior to the camp necessary permissions from the University authorities was procured.

The Principal of Ethel School was informed about the aims and objectives of this Pilot study. In addition, the parents, school children, teachers and the Principal were made aware of the advantages of this study by the expert dental practitioner who in turn gave their oral consent as it was non-invasive to be part of this pilot study. The samples were individually coded with unique numbers and methodology was explained by the researcher, and the swabs were collected during the dental examination itself. All the children affected with caries were only included in this study. General oral hygiene check was done using standard instruments by the expert.

The dental expert examined the teeth very meticulously noting the decay in systematic dental terminology. A premeditated personal questionnaire (PPQ) on the basis of their age, gender, hygiene and habits which were associated with their present decayed tooth condition or dental caries this was prepared in accordance with WHO Oral health assessment [14]. Children were categorized into two groups Age 1–6 and 7–12. The other group was the adolescent 13–15 year olds. Simple analysis in the form of percentage calculation was done to understand the rampant nature and status of the disease. All equipment and apparatus used were treated with utmost care and precision. Every sample was administered tests in less than 12 h of sample collection.

## **2.4 Processing of Samples**

The samples were processed within 12 h of collection and inoculated into Todd-Hewitt broth tubes and incubated under anaerobic condition (10% H<sub>2</sub>, 10% CO<sub>2</sub> and 80% N<sub>2</sub>, AnaeroPacks, HiMedia, Mumbai) for 48–72 h at 37 °C.

### **2.4.1 Aciduricity of Samples**

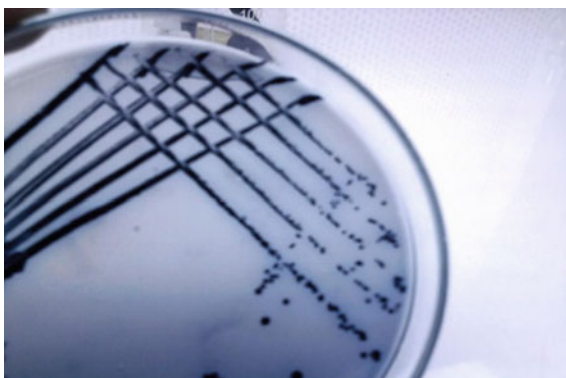
The samples if any showed a notable colour change in the Phosphate Buffer Saline, red to yellow, that sample was streaked directly on Mitis Salivarius Agar and incubated anaerobically (AnaeroPacks) for 48–72 h at 37 °C (Figs. 2 and 3).



**Fig. 2** *Streptococcus spp.*  
Visible polysaccharide  
production



**Fig. 3** Growth on Mitis  
Salivarius Bacitracin Agar  
(MSBA)



## 2.5 Isolation of Samples

### Media Preparation

The working bacteriological media were formulated and prepared as per earlier given protocols [16]. The plates prepared and poured were wrapped and stored at 4 °C.

### Identification of *Streptococcus mutans*

All the presumptive characteristics were confirmed using traditional culturing and isolation techniques (Tables 1, 2 and 3) as given in Bergey's Manual of Determinative Bacteriology [17].

### Morphological characteristics

The differential staining technique by Gram was fundamental to the phenotypic characterization of any bacteria.

### Cultural characteristics

The more conspicuous colonies were selected for further investigation [18, 19].

**Table 2** The incidence of children and adolescents in this study (3–19)

Age Group (In Years)		No. of Patients	No. of Patients Positive for <i>Streptococci</i> spp. (%)	Isolated <i>Streptococcus</i> spp. (%)	
				<i>Streptococcus mutans</i>	Other <i>Streptococcus spp.</i>
Children	1 – 6	23	09 (39.13)	09 (100)	00 (0)
	7 – 12	63	33 (52.38)	25 (75.76)	08 (24.24)
Adolescent	13 - 19	29	05 (17.24)	02 (40)	03 (60)
<b>Adults</b>					
<i>Youth</i>	20 – 45	55	14 (25.45)	09 (64.29)	05 (35.71)
<i>Middle-Aged</i>	46 – 60	21	11 (52.38)	08 (72.73)	03 (27.27)
<i>Senior Citizen</i>	≥ 60	11	05 (45.45)	03 (60)	02 (40)

## 2.6 Biochemical Characteristics

The isolate was tested for Catalase Activity, Voges-Proskauer Test and Hemolytic Activity (Fig. 4), one of the most important methods to identify the streptococcal species. Another test, Bile-Esculin Hydrolysis is used primarily for the presumptive identification and the differential isolation of esculin hydrolyzing Lancefield group D streptococci. The isolates from the samples gave an inconclusive result therefore the test was repeated using Esculin Agar without Bile. And finally, the acid producing capability was tested by using carbohydrates.

## 3 Results

India's economy and living style is stricken by unassuming and startling socio-economic and even political backgrounds thereby less than 90% only have visited a doctor for any dental related problem. The ruling party always hopes to target the children of socially compromised population and do many surveys to emancipate the poor standards of life in India [20].

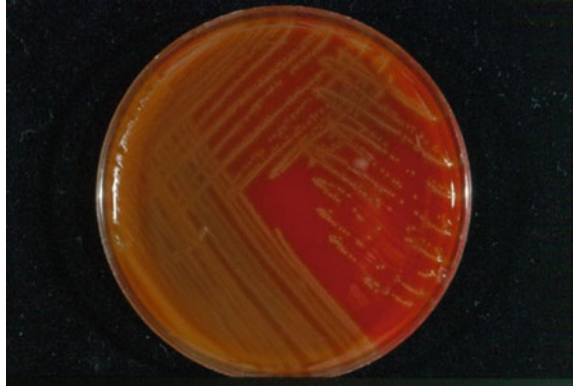
As the disease under study is a universal disease occurring in both genders, the above data reveals a very insignificant difference in the ratio of all the people involved in this study. Thus, ascertaining the fact that it is not biased when it comes to pathogenicity. Among the data obtained the total distribution of the microflora (Table 2). Along with the other *Streptococcus* spp. (9.82%), the desired organism *Streptococcus mutans* (33.82%) was also obtained (Table 3). Among the

**Table 3** Biochemical characteristics

S. No	INVESTIGATION	OBSERVATIONS	
		<i>Streptococcus mutans</i>	Other <i>Streptococcus spp.</i>
<b>1.0</b>	<b>CULTURE COLONY CHARACTERISTICS (Enhanced Anaerobic Condition)*<sup>1</sup></b>		
1.2	Colour of Colony on MSA <sup>A</sup>	Grayish Blue/Gray	Grayish Blue/Gray
1.3	Colour of Colony on SBA <sup>B</sup>	Gray	Gray
1.4	Colour of Colony on CA <sup>C</sup>	NT	NT
1.5	Colour of Colony on BA <sup>D</sup>	Grayish Brown	Grayish Brown
1.6	Colour of Colony on MSBA <sup>E</sup>	Grayish Blue	Gray/Blue
1.7	Form of the Colony	Irregular/Rough/Smooth	Irregular/Smooth
1.8	Elevation of the Colony	Convex	Convex/Pulvinate
1.9	Margin of the Colony	Undulate	Entire/Undulate
<b>2.0</b>	<b>MORPHOLOGICAL CHARACTERISTICS</b>		
2.1	Gram's Reaction	+ve	+ve
2.2	Shape of Cells* <sup>2</sup>	Oval	Oval/Spherical
2.3	Arrangement of Cells* <sup>2</sup>	Cocci in Pairs or Long , Short Chains	Cocci in Pairs or in Long or Short Chains
<b>3.0</b>	<b>BIOCHEMICAL CHARACTERISTICS</b>		
3.1	Hemolysis	$\alpha$	$\alpha$
3.2	Bile-Esculin Hydrolysis	-ve	+/-ve
3.3	Esculin Hydrolysis	+ve	+ve
3.4	Arginine Hydrolysis	-ve	-ve
3.5	Catalase Activity	-ve	-ve
3.6	VP* <sup>3</sup>	+ve	+ve
3.7	Acetoin Production	+ve	+ve
<b>4.0</b>	<b>CARBOHYDRATE FERMENTATION</b>		
4.1	Glucose	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.2	Sucrose	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.3	Mannitol	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.4	Sorbitol	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.5	Raffinose	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.6	Melibiose	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.7	Inulin	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
4.8	Trehalose	A <sup>+</sup> G <sup>-</sup>	A <sup>+</sup> G <sup>-</sup>
<b>5.0</b>	<b>VIRULENCE CHARACTERISTIC</b>		
5.1	Polysaccharide Detection	Dx/Lx	Lx

\*1 - Anerobe Gas Pack -5% CO<sub>2</sub> (Approximately); A - Mitis-Salivarius Agar(MSA), B - Sucrose Blood Agar(SBA), C - Chocolate Agar(CA), D - Blood Agar(BA), E - Mitis-Salivarius Bacitracin Agar(MSBA); +ve - Positive reaction; -ve - Negative reaction ; NT: Not Tested; \*2 – Under Microscopic Field; Alpha( $\alpha$ ) – Greenish-Brown discoloration , Beta ( $\beta$ ) - Complete Hemolysis and Gamma( $\gamma$ ) - Partial or No Hemolysis(No color change is seen); \*3 - Voges-Proskauer; Dx – Glucan, Lx – Levan.

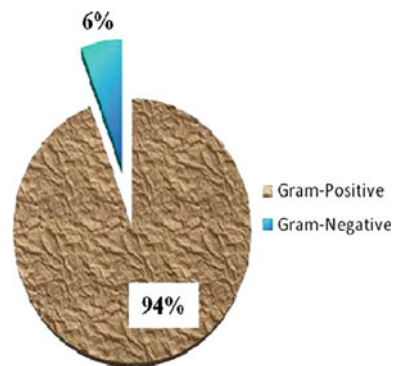
**Fig. 4** Hemolysis on blood agar (BA)



total 275 samples taken, 259 (94.18%) and 16 (5.82%) showed gram-positive and gram-negative microorganisms respectively (Fig. 5; Table 4).

The highest incidence of *Streptococcus mutans* was observed among Children (75-100%). The children and adolescent distribution is also noteworthy to observe that the Children in the age group 7–12, which is referred to as the mixed dentition period being the most affected age group (Fig. 1). Of the 114 positive patients analyzed, Streptococcal species were isolated in 66 (57.89%) patients (Table 1). Among them, *Streptococcus mutans* was isolated from 53 (80.30%) patients and *Streptococcus* spp. (other than *Streptococcus mutans*) was isolated from the rest 13 (19.69%) patients. Overall, among the children and adolescent group 80% of the incidence is by the principle etiology.

**Fig. 5** Total distribution of gram-positive and gram-negative microflora isolated from patients with dental caries



**Table 4** The total distribution of microflora

	No. of Samples	No. of Positive Patients	Micro flora Isolated (%)											
			Gram-Positive (94.18 %)								Gram-Negative (5.82 %)			
			<i>Streptococcus mutans</i>	Other <i>Streptococcus</i> spp.	<i>Micrococcus</i> spp.	<i>Enterococcus</i> spp.	<i>Bacillus</i> spp.	<i>Candida</i> sp.	<i>Staphylococcus</i> sp.	<i>Lactobacillus</i> spp.	<i>Eubacterium</i> spp.	<i>Fusobacterium</i> sp.	<i>Streptobacillus</i> sp.	<i>Neisseria</i> sp.
275	202	93 (33.82)	27 (9.82)	08 (2.91)	06 (2.18)	16 (5.82)	21 (7.64)	39 (14.18)	31 (11.27)	18 (6.55)	09 (3.27)	02 (0.73)	03 (1.09)	02 (0.40)

## 4 Discussion

India is a very unique country due to its variety of dietary habits, culture and huge population and the people of India follow different dietary lifestyles also due to religious and cultural reasons. The relevance of the present research lies on the reality that the perception of the many available indicators and factors of the disease makes possible the recognition of the vulnerable individuals.

The purpose of this study was to look at the influence of various factors involved for dental caries or decay that could be understood and evaluated. The authenticity of the need for such a research is evident in the diversified views recorded in different parts of the world.

Being a pioneer study, a total of 114 samples were taken from 115 patients. No decayed tooth was sampled twice. From each patient one sample was taken, as a unique separate sample. Another sample from the same patient was taken only and unless there was neighboring tooth affected with the same disease.

Among the total 275 samples taken, 259 (94.18%) and 16 (5.82%) showed gram-positive and gram-negative microorganisms respectively. Of the gram-positive organisms, the distribution was as follows: *Candida* sp. (5.82%), *Lactobacillus* spp. (11.27%), *Eubacterium* spp. (6.55%), *Staphylococcus* sp. (14.18%), *Bacillus* spp. (5.82%), *Enterococcus* spp. (2.18%), *Micrococcus* spp. (2.91%), other *Streptococcus* spp. (9.82%), and the desired organism *Streptococcus mutans* (33.82%). The distribution of the gram-negative organisms was as follows: *Veillonella* sp. (0.42%), *Neisseria* sp. (1.09%), *Streptobacillus* sp. (0.73%), and *Fusobacterium* sp. (3.27%) (Table 4). The lifestyles of bountiful enjoyed by developed countries and the livelihood of paucity not much enjoyed by developing or underdeveloped countries throws the disease into complete uncertainty of a decent and proper research for preventive recommendations whatsoever. The fact of the matter is that, the dietary habits and oral hygiene defines whether or not a particular strain of *Streptococcus mutans* would

become pathogenic. Prevalence of dental caries is reported to be about 50–60% and very few Indian studies have been carried out. *Streptococcus mutans* displays extraordinary compliance for sugar metabolism. WHO standards calls for a possibility to examine 2 or 3 classes of 12-year-olds of different socio-economic levels, in 2 or 3 local, easily accessible schools, where the widest possible differences in disease may be expected. If more than 20% of the children in the class are caries-free, the caries prevalence is low; if 5–20% is caries-free, the prevalence is moderate; and if fewer than 5% are caries-free, the prevalence is high [14]. Diverse Religious and Cultural lifestyles of India make Oral diseases a major public health crisis. Microbiological analysis of *Streptococcus mutans* among the patients living in sub-standard living and hygienic conditions proves to be a vital factor for incidence of Dental caries. Moreover, on a global scale, over the past decades, WHO has encouraged Member States to report information on disease level for making international comparisons? Major focus being on the dentition status, prosthetic status and needs, dental caries and dental treatment needs etc. Very ambitious goals have been formulated for oral health to be achieved by the year 2020 [21].

**Conflict of Interest** I declare that there is NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript and there is NO conflict of Interest.

## References

1. I.R. Hamilton, E.J.S. Martin, Evidence for the involvement of proton motive force in the transport of Glucose by a Mutant of *Streptococcus mutans*, Strain DR0001 defective in Glucose-Phosphoenolpyruvate Phosphotransferase activity. *Infect. Immun.* **36**(2), 567–575 (1982)
2. H.C. Slavkin, *Streptococcus mutans*, early childhood caries and new opportunities. *J. Am. Dent. Assoc.* **130**, 1787–1792 (1999)
3. M.R. Becker, B.J. Paster, Molecular analysis of bacterial species associated with childhood caries. *J. Clin. Microbiol.* **40**(3), 1001–1009 (2002)
4. J.A. Aas, B.J. Paster, L.N. Stokes, I. Olsen, F.E. Dewhirst, Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* **43**(11), 5721–5732 (2005)
5. K. McNeill, I.R. Hamilton, Acid tolerance response of biofilm cells of *Streptococcus mutans*. *FEMS Microbiol. Lett.* **221**, 25–30 (2003)
6. K. McNeil, I.R. Hamilton, Effects of acid stress on the physiology of biofilm cells of streptococcus mutans. *Microbiology* **150**, 735–742 (2004)
7. D.D. Zinner, J.M. Jablon, A.D. Aran, M.S. Saslaw, Experimental caries induced in animals by streptococci of human origin. *Proc. Soc. Exp. Biol. Med.* **118**, 766–770 (1965)
8. D. Brathall, Demonstration of five serological groups of streptococcal strains resembling *Streptococcus mutans*. *Odont. Rev.* **21**, 143–152 (1970)
9. D. Beighton, R.R. Russell, H. Hayday, The isolation and characterization of *Streptococcus mutans* serotype-*h* from dental plaque of monkeys (*Macaca fascicularis*). *J. Gen. Microbiol.* **124**, 271–279 (1981)

10. C.E. Safford, J.M. Sherman, H.M. Hodge, *Streptococcus salivarius*. J. Bacteriol. **33**(3), 263 (1937)
11. C.V. Loveren, J.F. Buijs, Similarity of Bacteriocin activity profiles of *Mutans streptococci* within the Family, when the children acquire the strains after the age of 5. Caries Res. **34**(6), 481–485 (2000)
12. J.M. Tanzer, J. Livingston, The microbiology of primary dental caries in humans. J. Dent. Edu. **65**(10), 1028–1037 (2001)
13. Y. Li, P.W. Caufield, Mode of delivery and other maternal factors influence the acquisition of *Streptococcus mutans* in infants. J. Dent. Res. **84**(9), 806–811 (2005)
14. WHO, *Oral Health Surveys Basic Method*, 4th edn. (World Health Organization, Geneva, 1989), pp. 760–871
15. A.K. Wan, W. Seow, L.J. Walsh, P.S. Bird, Comparison of five selective media for the growth and enumeration of *Streptococcus mutans*. Aus. Dent. J. **47**(1), 21–26 (2002)
16. O.G. Gold, H.V. Jordan, J. Van Houte, A selective medium for *Streptococcus mutans*. Arch. Oral Biol. **20**, 473–477 (1973)
17. J.K. Clarke, On the bacterial factor in the etiology of dental caries. Br. J. Exp. Path. **5**(3), 141–147 (1924)
18. S. Edwardsson, Characteristics of caries-inducing human streptococci resembling *Streptococcus mutans*. Arch. Oral Biol. **13**, 637–646 (1968)
19. M. Seki, Y. Yamashita, Y. Shibata, H. Torigoe, H. Tsuda, M. Maeno, Effect of mixed *Mutans streptococci* colonization in caries development. Oral. Microbiol. Immun. **21**, 47–52 (2006)
20. Oral Survey Report, in *Survey Indicates Poor Standards of Oral Health in India* (2010). Available from <http://www.imrbint.com/downloads/media-room/OralSurvey>. Accessed 1 Oct 2017
21. M.P. Hobdell, P.E. Petersen, J. Clarkson, N. Johnson, Global goals for oral health by the year 2020. Int. Dent. J. **53**, 285–288 (2003)

# Screening of Genetic Variance Based on CO-I Gene Analysis of Silkworm (*Bombyx mori*) Races



S. Vimala, Sriramadasu Kalpana, EI-Sheikh A. EI-Syed,  
and D. M. Mamatha

**Abstract** Genetic variability of 29 domesticated races of mulberry silkworms were collected from different authorized silkworm germplasm centers of South India were studied based on the sequences of mitochondrial cytochrome c oxidase I (COI) gene. Partial COI gene regions of 29 specimens were amplified and sequenced. All the sequences had more or less similar COI gene sequence information and are having strong bias towards higher ‘A’, ‘T’ contents, more translational substitutes. Neighbor-joining, maximum likelihood and Bayesian methods were used for studying the phylogenetics of Intraspecific relationships among 29 specimens. The above study forms a basis to develop innovative measures and strategies for the conservation of natural diversity existing among these distinct Silkworm races. This knowledge will also help in recognizing significant breeds of mulberry silkworm, which leads a way to identify promising potential lines to design successful breeding programs to enhance silk productivity and resistance against diseases and temperatures.

**Keywords** Intraspecies · Mitochondrial cytochrome c oxidase I (COI) · Silkworm germ plasm · Genetic variance

---

S. Vimala · S. Kalpana · D. M. Mamatha (✉)

Molecular Cloning Lab, Department of Biosciences and Sericulture, Sri Padmavati Women’s University, Tirupati, AP, India

e-mail: [dmmfulbrightucdavis@gmail.com](mailto:dmmfulbrightucdavis@gmail.com)

S. Vimala

e-mail: [emerald.leena7@gmail.com](mailto:emerald.leena7@gmail.com)

S. Kalpana

e-mail: [sriramadasu.kalpana@gmail.com](mailto:sriramadasu.kalpana@gmail.com)

EI-Sheikh A. EI-Syed

Plant Protection Department, Zagazig University, Zagazig, Egypt

e-mail: [eaelsheikh@agri.zu.edu.eg](mailto:eaelsheikh@agri.zu.edu.eg)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,

Learning and Analytics in Intelligent Systems 15,

[https://doi.org/10.1007/978-3-030-46939-9\\_25](https://doi.org/10.1007/978-3-030-46939-9_25)



## 1 Introduction

The similarities and differences in morphological features, quantitative and qualitative traits have been used to group and classify the organisms. A group wise classification of the organisms has been made based on the external features like shape, structure and form. Changes in the environmental factors may cause minor differences among morphs, races, breeds, strains, biotypes and populations. Due to their high polymorphism and independent nature to environmental factors, DNA based molecular markers plays a major role in vast areas of biology which includes evolution, ecology, phylogenetic studies, population genetics and population dynamics in plants and animals [2]. Nucleotide substitution variations may involve in attaining details about phylogenetic relationship and structure of population community of different organisms. Those variations causes increase in diversification among independent lineages and divergent taxa [6] and also show slight changes at taxonomic level. Mitochondrial cytochrome c oxidase I (COI) gene is most popularly used marker to study the molecular systematics. Most of the research studies are focused on this gene which contains~658 bps to identify as unique code for many species [5, 19]. Quick and precise identification of species can be done using mitochondrial COI gene [22]. In addition to that COI gene is also used to recognize new species. Research studies revealed that mitochondrial COI gene could effectively be used as DNA barcodes which was known to be a common tool to identify and discover new species in a accurate way and as well as assessing biological diversity based on molecular data [7, 23]. Mulberry silkworm *Bombyx mori* is an economically important, lepidopteran model insect belonging to the Bombycidae family [6]. Silkworms are typically evolved phenotypes over a short history of domestication and evidence for remarkable phenotypic changes. According to Xia et al. [25] genetic divergence mechanism between two species is significant for the evolution. By using different molecular markers, better understanding of intraspecific diversity and polymorphism in silkworms is possible [3, 4, 17].

Estimation of genetic variations and phylogenetic relationship between and within species are compulsory requirements to be able to maintain at a certain level. Information on genetic variation differences among different populations of a single species can be exploited in breeding programmes. Different molecular markers have been used to resolve intraspecific biodiversity and polymorphism in silkworms which aids in designing better breeding programs. Majority of the research have been carried out to determine the genetic variability of silkworms. However *B. mori* is a distinct silk producing species in Bombycidae family. Very less genetic variability was seen within these species. Construction of phylogenetic tree is a very challenging task in closely related species because of existence of subtle genetic divergence. Hence it is essential to identify the genetic variance between races of the mulberry silkworm which made a way to develop better breeding programs. The current study involves the determination of phylogenetic relationships, and intraspecific distances based on the COI mitochondrial molecular marker referred as DNA barcodes among twenty nine local races and strains of mulberry silkworms collected from South India

germplasm centers. Here, we discussed the intraspecific distances among twenty nine races and strains of mulberry silkworm. The data generated in this study forms a basis for developing new strategies to design successful breeding programs for promising potential lines to enhance hybrid vigor, resistance against diseases and silk productivity.

## 2 Materials and Methods

### a. Sample collection:

Twenty nine races and strains of mulberry silkworm were collected from authenticated germplasm banks of APSSRDI, Hindupur and CSRTI, Mysore. Each individual sample was stored in sample container filled with absolute ethanol and stored in  $-20^{\circ}\text{C}$  freezer for further experiments.

### b. Genomic DNA isolation:

The mitochondrial cytochrome c oxidase I gene is a commonly used DNA marker for analyzing inter and intraspecific relationships of silkworms and other lepidopterans [1, 8]. Hence, the nucleotide sequences of mitochondrial COI gene was selected for the current study. For genetic analysis of mulberry silkworm races, Genomic DNA from individual larva of twenty nine races was isolated. The isolations were done by using Promega Genomic DNA isolation kit according to the manufacturer's instructions. Finally the DNA was eluted with Nuclease free water. The quality of extracted DNA was checked on 1% agarose gel electrophoresis. The concentration of genomic DNA was measured by using nano-spectrophotometer and diluted to  $50\text{ ng}/\mu\text{l}$  for PCR amplifications. All chemicals used in this experiment were molecular grade chemicals purchased from Sigma Chemicals company.

### c. Amplification and sequencing of Mitochondrial COI gene:

For the amplification of mitochondrial COI gene, primers specific to lepidopteran insects i.e., Lep F1—ATTCAACCAATCATAAAGATATTGG, Lep R1 (single space between Lep and R1)—TAAACT TCTGGATGTCCAAAAAATCA were acquired from BOLD Systems online data base [5] and used in PCR. Amplification was carried out in a Thermo-Cycler with  $15\ \mu\text{L}$  reaction mixture contained  $1.25\ \mu\text{l}$  of 10XPCR buffer  $8.3\ \mu\text{L}$  of Trehalose,  $0.4\ \mu\text{l}$  of  $15\ \text{mM}$   $\text{MgCl}_2$ ,  $1\ \mu\text{L}$  of  $2.5\ \text{mM}$  dNTP,  $1\ \mu\text{L}$  of  $10\ \mu\text{M}$  forward primer;  $1\ \mu\text{l}$  of  $10\ \mu\text{M}$  reverse primer,  $1\ \mu\text{L}$  of DNA ( $50\ \text{ng}/\mu\text{L}$ );  $0.1\ \mu\text{L}$  *Taq* DNA polymerase ( $5\ \text{U}/\mu\text{L}$ ) and  $0.95\ \mu\text{L}$  double distilled water. The PCR reaction conditions were 1 cycle of  $94^{\circ}\text{C}$  for 4 min, followed by 35 cycles of  $94^{\circ}\text{C}$  for 30 s (denaturation),  $44^{\circ}\text{C}$  for 1 min (annealing),  $72^{\circ}\text{C}$  for 3 min (elongation) and a final extension of 5 min at  $72^{\circ}\text{C}$ . The amplified PCR products were resolved on 1.5% Agarose gel electrophoresis. Further, successfully amplified COI gene products were purified enzymatically with EXOSAP (EXO SAP: EXO1— $0.25\ \mu\text{L}$ , SAP— $0.5\ \mu\text{L}$ , 10X SAP Buffer—1, Template— $2.5\ \mu\text{L}$ ) followed by cycle sequencing methods

prior to sequencing. Sequencing of both strands of DNA samples was done using Big Dye Terminator V.3.1 Cycle sequencing Kit in Capillary Sequencer (AB1 3130 Genetic analyzer).

**d. Sequence editing and data analysis:**

The obtained sequences of mitochondrial COI regions were validated and edited using Codon-code aligner. All the COI sequences of mitochondrial DNA regions were homologous in length and edited sequences could be easily assembled and aligned using ClustaX2 [21] program. BLAST search was performed to identify the similarity of the sequences. After alignment the ends were trimmed and sequences were submitted to BOLD database to obtain DNA barcodes. Analysis of the obtained sequences like nucleotide frequencies, nucleotide pair sequences and transition/transversion ratios, overall transition/transversion bias (R) and nucleotide substitutions per site were calculated by using maximum composite likelihood parameter in MEGA6 program [20]. All gaps and missing data were removed from dataset. DnaSP software [11] was used to identify the Insertion-deletion polymorphism (InDels), conserved, variable, parsimony informative and singleton sites. The parameters like pairwise nucleotide differences among DNA sequences (K), Nucleotide diversity (Pi), the number of nucleotides per site between two sequences, were determined to estimate the genetic diversity. All analysis was done by using DnaSP software. Intraspecific nucleotide distances were calculated with Kimura2 parameter method MEGA6 [20].

**e. Phylogeny studies:**

To explore phylogenetic relationship among the closely related races of *B. mori*, we constructed three phylogenetic trees based on the mitochondrial COI gene sequences. These three trees were generated using N-J method in MEGA6, Maximum likelihood criterion in MEGA6 and Bayesian method [9]. NJ (Neighbor-Joining) tree was generated using MEGA 6 [20] with Kimura-2-Parameter molecular evolutionary model. Statistical support for nodes and internodes were calculated using bootstrap analysis with 1000 replicates. The ambiguous sequence data was completely removed. Bayesian inference was calculated with the computer program MrBayes ver. 3.2.7 [16] with the best-fitting model general time reversible (GTR)  $\mu$ G for the evolution of mitochondrial sequences. The analysis was made by Markov Chain Monte Carlo (MCMC) algorithm for twice with two million generations, sampling the trees every hundredth generations with four independent chains running simultaneously and first 5000 trees were eliminated as burn-in. Posterior probabilities of the branching pattern were computed by the remaining trees with 50% majority-rule consensus tree. The results were visualized and checked using FigTree software package. Maximum likelihood phylogeny analysis was carried out in MEGA 6 adjusted to 1000 bootstrap replicates under Tamura 3-parameter (T92)  $\mu$ G model. The strongness of the clade was assessed using bootstrap analysis. In this study the generated phylogeny tree support values are mentioned as weak (40–60%), Moderate (61–75%), Good (76–88%) and strong support (>89%).

### 3 Results

The races collected from Germplasm center of APSSRDI, Hindupur were APS 72, APS 12, APS 71, APS 20, APSDR 105, APSHT 02, APS 33, APS 45 and the races collected from CSR&TI, Mysore were CSR 2, CSR 52, CSR 51, CSR 50, CT1PP, CSR 4, CSR 5, CSR 6, CSR 16, CSR 26, CSR 27, CSR 46, CSR 47, CSR 48, NB7, Kolar Gold, Hosa Mysore, Nistari, NB4D2, Kalimpong A, Pure Mysore. The sequencing results of PCR products of 29 races of *Bombyx mori* have shown 658 bps of COI gene. The obtained sequences were edited and have been submitted to BOLD database. The DNA barcodes were developed to all 29 sequences. Further all these 29 COI gene sequences were analyzed for genetic diversity studies.

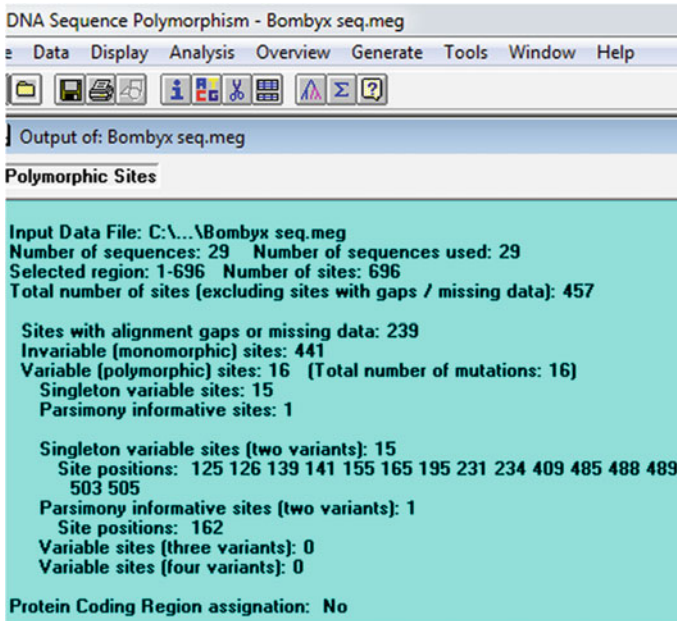
#### a. Genetic Diversity Analysis:

The genetic divergence among the races of mulberry silkworm were checked based on the sequence information of the mitochondrial locus COI gene. The total length of newly obtained COI gene region ranged between 619 to 658 bps respectively. A ~ 658 bps of COI gene sequences for all samples of mulberry silkworm races was analyzed. The differences among the sequences due to nucleotide substitutes which were called as In-Dels (Insertion-Deletion polymorphism) are not seen in COI gene sequences. In the total selected region of 696 bps, 441 bps were conserved sites, 16 bps were variable sites and 239 bps were alignment gaps and missing data. The 16 bps of variable sites includes 15 singleton and 1 parsimony informative sites (Fig. 1). High content of 'A' and 'T' nucleotides were seen in COI gene sequences. The average frequencies of T, C, A and G nucleotides of COI gene sequences were 36.7, 15.5, 33.5 and 14.3% (Table 1). The rate of interchange of purines or pyrimidines i.e., translational substitutes are more when compared to the rate of interchange between purines and pyrimidines i.e., transversional substitutes in COI gene sequences. The overall transition/transversion ratio  $R = 0.982\%$ .

After pair-wise genetic divergence analysis, it was found that the sequences of APS 71, APSDR 105, APS 45, APS 72, APSHT 02, CSR 4, CSR 5, CSR 6, CSR 16, CSR 26, CSR 27, CSR 46, CSR 47, CSR 48, CT1PP, Hosa Mysore, Kalimpong A, Kolar Gold, NB4D2, NB7, Nistari, Pure Mysore were near to each other. Overall estimation of mean diversity calculations between the sequences (d) was 0.003 (Fig. 2). Lowest genetic divergence was found between the sequences of APS 72, APS 71, APS 12, CSR2, CSR52, CSR51. The COI gene sequences of APS12, CSR2, CSR52 and CSR51 has shown variable genetic divergence with remaining races. The highest genetic distance was found between APS20 and CSR 50 (0.018). Average number of nucleotide differences per site between two COI gene sequences ( $P_i$ ) was 0.00268% and average number of pairwise nucleotide difference (K) was 1.227%.

#### b. Phylogenetic analysis:

Molecular phylogenetic trees based on mitochondrial partial COI gene sequences were constructed using NJ, ML and BI methods showed more or less identical clustering. In this study, phylogeny studies were done within races of the same species. Hence there was no out group for the formation of trees. In NJ phylogenetic tree there



**Fig. 1** Details of polymorphism of COI gene sequences among twenty nine races of *Bombyx mori*

was no branch length (0.000) difference in the races of CSR 26, Pure Mysore, Kolar Gold, APS71, NB7, Kalimpong A, CSR 6, APSDR 105, Hosa Mysore, CSR 4, CSR 27, CSR 47, APSHT02, APS 72, APS 45, NB4D2, CSR 16, CT1PP, Nistari, CSR 5, CSR 46, CSR 48. These results revealed that the above races doesn't shown any variance in their COI gene sequence information. CSR 51, CSR52, CSR 2 having same branch length 0.002 a little bit more than the above races and strains. APS 20, APS 12 and APS33 shared as sister clades and having different branch lengths i.e., 0.004, 0.000, 0.011. But CSR 50 and APS 33 showed high branch length (0.011) than remaining races. The above results were shown in Fig. 3.

Likewise NJ tree, In ML tree (Fig. 4) CSR 50, CSR 52, CSR 51, APS 20 and APS 33 showed variable branch length among twenty nine races of mulberry silkworms. CSR 50 and APS 33 revealed higher branch length. APS 12, APS 20 and APS 33 shared as sister clades. Remaining all races of mulberry silkworms has similar branch length.

In BI tree (Fig. 5) CSR 51, CSR 50, APS 33 and CSR 52 forms paraphyly group where CSR 50, APS 33 has same branch length. But CSR 51 and CSR 52 indicates difference in their branch lengths. Remaining races forms separate clusters and showed variation in their branch lengths.

**Table 1** The average nucleotide frequencies of partial COI gene sequences of twenty nine races of mulberry silkworm

	T(U)	C	A	G	Total
APS72	36.5	15.9	33.8	13.8	616.0
APS12	36.5	15.8	33.6	14.1	622.0
APS71	37.2	15.2	33.6	14.0	658.0
CSR2	36.6	16.1	33.5	13.7	483.0
APSDR105	36.0	15.4	34.0	14.6	583.0
APS20	36.4	16.2	33.1	14.3	544.0
APSHT02	36.6	15.5	33.4	14.4	631.0
CSR52	35.6	15.7	33.9	14.8	610.0
CSR51	36.5	15.7	33.5	14.3	561.0
CSR50	35.9	15.5	34.2	14.5	608.0
CT1PP	36.3	15.8	33.2	14.8	609.0
APS33	36.1	16.4	32.8	14.7	646.0
APS45	35.5	15.9	34.2	14.4	605.0
NB7	37.0	15.4	33.3	14.2	648.0
Kolar Gold	37.1	15.3	33.6	14.0	542.0
Hosa Mysore	37.5	15.0	33.7	13.9	606.0
Nistari	37.2	15.3	33.3	14.2	648.0
NB4D2	37.2	15.3	33.3	14.2	648.0
Kalimpong A	36.9	15.3	33.3	14.5	636.0
Pure Mysore	37.2	15.3	33.3	14.2	648.0
CSR 4	37.2	15.3	33.3	14.2	648.0
CSR 5	36.9	15.4	33.4	14.3	643.0
CSR 6	36.3	15.7	33.9	14.1	631.0
CSR 16	37.2	15.3	33.3	14.2	648.0
CSR 26	37.2	15.3	33.3	14.2	648.0
CSR 27	37.2	15.3	33.3	14.2	648.0
CSR 46	37.2	15.3	33.3	14.2	648.0
CSR 47	37.2	15.3	33.3	14.2	648.0
CSR 48	36.9	15.3	33.5	14.3	645.0
Avg.	36.7	15.5	33.5	14.3	619.3

## 4 Discussion

The COI gene region is one of the best standard method has been used as species level identification in insects. The generated phylogenetic trees based on DNA sequences in the current study are allowed to clustered the closely related and separate the variables in Silkworm races. In this study, COI showed the high rate of amplification

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1. APS12																												
2. APS20	0.004																											
3. APS33	0.011	0.016																										
4. APS45	0.002	0.007	0.013																									
5. APS71	0.002	0.007	0.013	0.000																								
6. APS72	0.002	0.007	0.013	0.000	0.000																							
7. APSDR105	0.002	0.007	0.013	0.000	0.000	0.000																						
8. APSHT02	0.002	0.007	0.013	0.000	0.000	0.000	0.000																					
9. CSR 19B	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000																				
10. CSR 11A	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000																			
11. CSR 12A	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000																		
12. CSR 13B	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000																	
13. CSR 14A	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000																
14. CSR 15A	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000															
15. CSR 16	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000														
16. CSR 17	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000													
17. CSR 9A	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000												
18. CSR2	0.004	0.009	0.016	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002										
19. CSR30	0.013	0.018	0.025	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011									
20. CSR51	0.004	0.009	0.016	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.013							
21. CSR52	0.004	0.009	0.016	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.013	0.004						
22. CT1PP	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000					
23. Hosa Mysore	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000				
24. Kalimpong A	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000	0.000			
25. Kolar Gold	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000	0.000	0.000		
26. NB4D2	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000	0.000	0.000	0.000	
27. NB7	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000	0.000	0.000	0.000	
28. Nistari	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000	0.000	0.000	0.000	
29. Pure Mysore	0.002	0.007	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002	0.002	0.000	0.000	0.000	0.000	0.000	

Fig. 2 Pairwise genetic distances of races of mulberry silkworm based on mitochondrial COI partial gene sequences

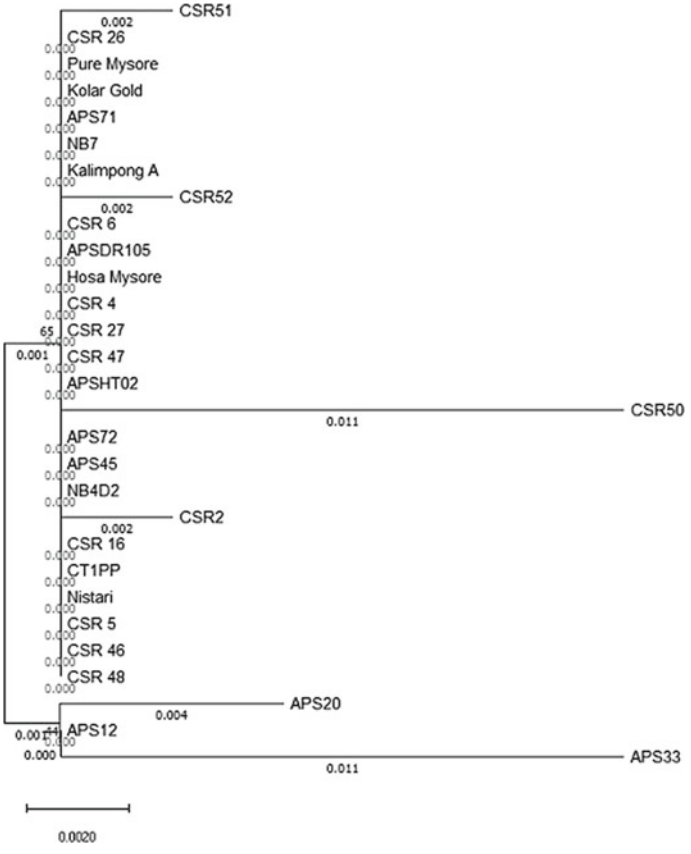
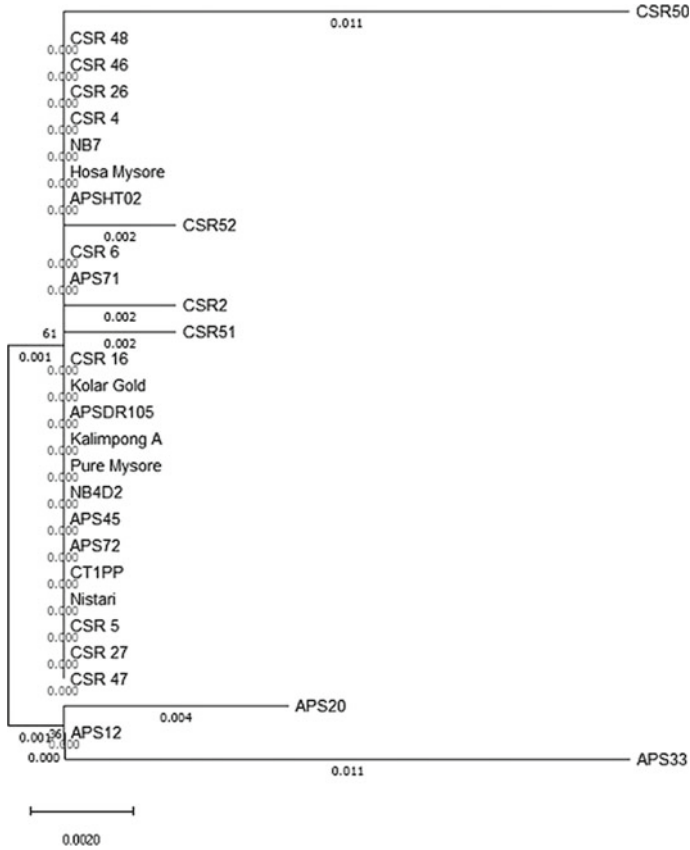


Fig. 3 Phylogenetic tree explaining distances among twenty nine races of *Bombyx mori* species based on mitochondrial partial COI gene sequences by Neighbor-Joining method. The bootstrap values are shown at the branching points



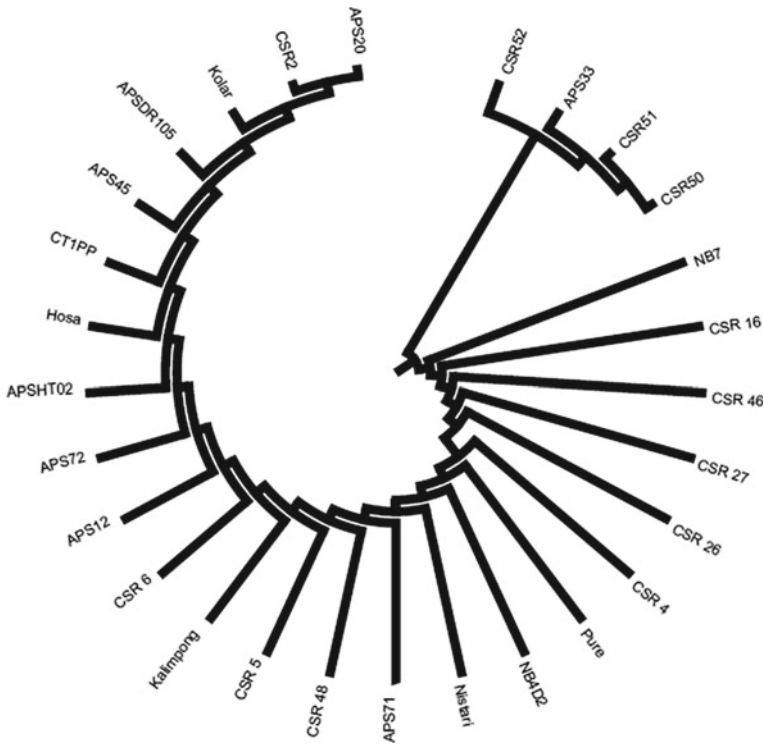
**Fig. 4** Phylogenetic tree explaining distances among twenty nine races of *Bombyx mori* species based on mitochondrial partial COI gene sequences by maximum likelihood criterion. The bootstrap values and branch length were shown

and sequencing in all of the specimens tested. The phylogeny studies within races of the species *Bombyx mori* were resolute by the NJ, ML and BI methods using mitochondrial COI gene sequences. This study is the first inclusive phylogenetic analysis of the evolutionary relationships within races of *B.mori* species based on molecular data.

Mitochondrial COI sequences for 29 races belonging to *B.mori* species were submitted to BOLD data base for generating DNA barcodes. The COI gene in mitochondrial genome has been proved as popular tool for the identification of closely related species of the family Bombycidae.

However, some studies showed that twelve races of *B. mori* were extensively distributed worldwide. The sequence variance among them was very small (0–0.2%) [10]. Utilization of mitochondrial gene in particular has been difficult in some cases because most of the sequences were rich in high A and T in insects [24]. Hence,





**Fig. 5** Phylogenetic tree explaining distances among twenty races of *Bombyx mori* species based on mitochondrial partial COI gene sequences by Bayesian inference method

**Table 2** Rates of different transitional substitutions are **(bold)** and those of transversional substitutions (*italics*) in partial COI gene sequences of mulberry silkworm

	A	T/U	C	G
A	–	8.3	3.67	<b>10.85</b>
T/U	7.61	–	<b>5.13</b>	3.03
C	7.61	<b>11.61</b>	–	3.03
G	<b>27.2</b>	8.3	3.67	–

identified results within the clades of the generated tree were supported by the COI gene region. The three phylogenetic trees showed slight difference in tree topologies between COI gene sequences. Molecular marker studies lead a way to study the stability and variability of genotypes. These studies helping the breeders thrive in breeding different new varieties which show evidence that these markers make significant role in conservation of Silkworm races [18]. According to the pair wise gene distances, the lowest distance of 0.002 was found between APS 72, APS 71, APS 12, CSR 2, CSR 52 and CSR 51. The alignment of the COI gene sequences of

the above races showed low diversity without addition or deletions of nucleotide base pairs. Generally, species delimitation is calculated by the gap between intraspecific and interspecific divergences in various animal groups [8, 12–15]. The phylogenetic trees constructed by NJ, ML and BI methods indicated almost similar topology. Phylogenetic studies results showed that the APS 20, APS 12, APS 33 races forms paraphyl group with weak nodal support in NJ and ML tress. But in BI tree CSR 51, CSR 50, APS 33 and CSR 2 formed as paraphyl group. In the present study CSR 51, CSR52, CSR 50, CSR 2, APS 12, APS 33 races had a distant relationship with remaining races. Comparatively unambiguous relationships of the races within *B. mori* were presented for the first time. Based on the obtained results of the current study, further research on molecular analysis can be done for silkworm races which show high divergence can be used as better parental stock to develop the desired lines.

## 5 Conclusion

Based on the current research study we concluded from the results of three generated phylogenetic trees, CSR 51, CSR 52, CSR 50, APS 20 and APS 33 races of silkworm *Bombyx mori* showed high divergence among the other. Hence, the above breeds can be explored as better parental lines in breeding programs to develop better performing breeds for productive silk industry.

## References

1. K.P. Arunkumar, M. Metta, J. Nagaraju, Molecular phylogeny of silkmoths reveals the origin of domesticated silk moth, *Bombyx mori* from Chinese *Bombyx mandarina* and paternal inheritance of *Antheraea proylei* mitochondrial DNA. *Mol. Phylogenet. Evol.* **40**, 419–427 (2006)
2. S.K. Behura, Molecular marker systems in insects: current trends and future avenues. *Mol. Ecol.* **15**, 3087–3113 (2006)
3. S.N. Chatterjee, R.K. Datta, Hierarchical clustering of 54 races and strain of mulberry silkworm, *Bombyx mori*: Significance of biochemical parameters. *Theor. Appl. Genet.* **85**, 394–402 (1992)
4. M. Eguchi, Alkaline phosphatase isozymes in insects and comparison with mammalian enzyme. *Comp. Biochem. Physiol.* **111**, 151–162 (1995)
5. O. Folmer, M. Black, W. Hoeh, R. Lutz et al., DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**, 294–299 (1994)
6. N. Galtier, D. Enard, Y. Radondy, E. Bazin et al., Mutation hot spots in mammalian mitochondrial DNA. *Genome Res.* **16**, 215–222 (2005)
7. P.D.N. Hebert, A. Cywinska, S.L. Ball, J.R. deWaard, Biological identifications through DNA barcodes. *Proc. Biol. Sci.* **270**, 313–321 (2003)
8. P.D.N. Hebert, E.H. Penton, J.M. Burns et al., Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl. Acad. Sci.* **101**, 14812–14817 (2004)

9. J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001)
10. Y.L. Jiang, *The origination and Differentiation of Domesticated Silkworms* (Jiangsu Scientific and Technical Press, Nanjing, 1982)
11. P. Librado, J. Rozas, DnaSP v5: software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009)
12. R. Meier, S. Kwong, G. Vaidya et al., DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.* **55**, 715–728 (2006)
13. R. Meier, G. Zhang, F. Ali, The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. *Syst. Biol.* **57**, 809–813 (2008)
14. C.P. Meyer, G. Paulay, DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* **3**, 422 (2005)
15. N. Puillandre, A. Lambert, S. Brouillet, ABGD, automated barcode gap discovery for primary species delimitation. *Mol. Ecol.* **2**, 1864–1877 (2012)
16. F. Ronquist, J.P. Huelsenbeck, MrBayes 3—Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003)
17. M. Salehi Nezhad, S.Z. Mirhosseini, S. Gharahveysi, M. Mavvajpour, A.R. Seidavi, Analysis of genetic divergence for classification of morphological and larval gain characteristics of peanut cocoon silkworm (*Bombyx mori L.*) *Germplasm. Am.-Eurasian J. Agri Environ. Sci.* **6**(5), 600–608 (2009)
18. S. Bakkappa, E. Talebi, G. Subramanya, Role of molecular markers (RAPD & ISSR) in silkworm conservation. *I.J.A.B.R.* **1**(1), 01–07 (2011). ISSN 2250-3579
19. M. Stoeckle, Taxonomy, DNA and the bar code of life. *Bioscience* **53**, 2–3 (2003)
20. K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013)
21. J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin et al., The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997)
22. S. Vimala, D.M. Mamatha, et al., DNA Barcoding studies of different mulberry silkworms and their phylogeny based on computational tools. *IJSCME—SCSMB-16 March-2016*. ISSN-2349-8439 (2016)
23. R.D. Ward, R. Hanner, P.D. Hebert, The campaign to DNA barcode all fishes. *FISH-BOL. J. Fish Biol.* **74**, 329–356 (2009)
24. T. Wirth, R. Le Guellec, M. Veuille, Directional substitution and evolution of nucleotide content in the cytochrome oxidase II gene in earwigs (Dermapteran insects). *Mol. Biol. Evol.* **16**, 1645–1653 (1999)
25. Q. Xia, Y. Guo, Z. Zhang, D. Li, Z. Xuan, Z. Li, F. Dai, Y. Li, D. Cheng, R. Li et al., Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**(5951), 433–436 (2009)

# A Collaborative Filtering Based Ranking Algorithm for Classifying and Ranking NEWS TOPICS Using Factors of Social Media



S. Gayathri Devi, K. R. Manjula, and K. Subhashri

**Abstract** Searching topic and tracking the headlines on Topic Detection and Tracking (TDT) makes the users more convenient to see what is happening in the real-world through internet. Due to the large quantity of news articles, it is not possible to see all the topics. It makes the necessity of topic ranking in terms of time and importance. Topic ranking is based on frequency of occurrence of the topic in media and the amount of attention paid by the users. Both these factors are time dependent. So it is necessary to include the effect of time. However, inconsistency always exists between these two factors. In this paper, an automatic online news topic ranking algorithm has been proposed. The analysis between Media Focus (MF) and User Attention (UA) has been carried out in the proposed algorithm in terms of inconsistency. Here, UI defines the strength of the community who discusses the topic. Overlapping Topic Clusters (TCs) is found by Hybrid Fuzzy Clustering (HFC) approach. Artificial Bee Colony Optimization (ABCO) is performed to calculate the node weight. It is necessary to personalize the SociRank to present different topics for different users. Collaborative Filtering based Ranking Algorithm (CFRA) is adopted for ranking. The proposed CFRA based SociRank improves the quality and variety of automatically identified news topics. The same has been proven by experimental results.

**Keywords** News topic detection and tracking · Topic ranking · Hybrid fuzzy clustering · Artificial bee colony optimization · Collaborative Filtering based Ranking Algorithm

---

S. Gayathri Devi · K. R. Manjula (✉) · K. Subhashri  
Department of CSE | School of Computing, SASTRA Deemed to be University, Thanjavur, India  
e-mail: [manjula@cse.sastra.edu](mailto:manjula@cse.sastra.edu); [manju\\_sakvarma@yahoo.co.in](mailto:manju_sakvarma@yahoo.co.in)

S. Gayathri Devi  
e-mail: [gayathridevi@cse.sastra.edu](mailto:gayathridevi@cse.sastra.edu)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_26](https://doi.org/10.1007/978-3-030-46939-9_26)

# 1 Introduction

Many web-based news applications such as google news, daily hunt etc., are used to organize a news topics. These news topics are created by the news stories gathered from various websites. Topic Detection and Tracking (TDT) [1] is used for automatic updation and construction of these news topics. TDT felicitate the users to know the current trends and happenings things around us. But, it is not possible for the user to view all the news topics due the amount of news topics. It is important to rank the news and prioritize them based on the importance.

News topics are ranked by using two different rules [2]. (i) Based on the timing of the topic update, (ii) based on the number of included news stories. First rule is used for getting timely results. Second rule is used to rank the stories based on the number of topics. Other than these simple rules, few more parameters should be considered for ranking the news topics. They are, contribution of the news importance and number of users that news topic attracts.

Media focus and user attention are two important basic factors used for ranking the news. Media focus defines the magnitude of the news topic reported by the websites. User attention defines the interest of the user in reading the news stories. Both the factors are function of a time. The effect of time in ranking the news by these parameters have already been defined. There exists a variation between these parameters, which is given as follows. Close attention of the users are not attracted by a topic reported frequently by the media. This is due to the wrong judgment of the user, news title, news subjects and positioning of the news.

So media focus itself is not enough for the topic ranking. Poorly reported topics may yield many users due to the titles or by surprise or due to the closeness to the users life or effect of anchor. News stories are taken from authority pages [3], which are pointed to by news index pages.

In addition, anchor texts also play an vital role in the decision of viewing, as viewers often uses hyperlinks to see the news stories. Attractive anchor texts may yield many user attention. So the single factor UA is not adequate for topic ranking.

The difference between media focus and user attention is examined and quantitative calculation of the same is carried out in this paper. CFRA is used to calculate these parameters automatically in online. Obviously, the news, which is focused by media and normal users attention are ranked high. The problems to be rectified during the news ranking are: (1) Calculation of media focus and user attention and difference between them. (2) How to account all the mentioned factors in ranking?

To effectively identify the news topics, an unsupervised system—SociRank—is proposed. This identifies news topics from both social media and the news media. The factors used for topics ranking are MF, UA, and UI. The proposed system is applicable in various fields.

Other features of the proposed SociRank is keyword extraction, measures of similarity, graph clustering, and social network analysis. This effective topic ranking is evaluated in both controlled and uncontrolled manner by doing experiments with given data set on the proposed system.

Overlap with social media is identified by the keywords from news media sources. A graph is plotted by using these keywords and their co-occurrences. Distinct topics are identified by the clustering of the graph. MF, UA, and UI are calculated after clustering. These three measuring factors are finally combined together to get best result.

In this paper, Sect. 2 explores the past inventions in the field of topic detection; Sect. 3 proposes the CFRA based SociRank for online news topics. Section 4 describes the experimental data and results whereas Sect. 5 summarizes the work with future enhancement.

## 2 Related Work

The merits and demerits of existing topic detection and ranking based algorithms are discussed in this section.

Fung et al., determined the bursty features in different time windows by using the time information [4]. It is a feature-pivot clustering approach. Bursty features are grouped to detect the bursty topics. Feature grouping is done based on the feature distribution. At last, peak periods of major topics are indexed.

Chen et al., used multi-dimensional sentence modeling and timeline analysis to extract the hot topic [5]. Mapping of distribution of terms is used to find the hot terms. Key sentences are identified and grouped as a cluster to find the hot topics. Popularity variation over time, appearance in news channels and news stories and strong continuity characteristics are used to find the hot topics.

He et al., used semi-automatic algorithm for topic detection [6]. “Incremental clustering algorithm and incremental TF-IDF model” is used for new topic detection. Topic report frequency and number of consecutive effective time units are used to rank the topics. The experimental results shows that this method is not well-suited.

More number of approaches have been developed for topic modeling in the past decade [7]. Additional structures are added in the topic model with the advancement in the MAP/ML estimate. MAP/ML estimates like variational Bayes or MCMC are used to get the efficient results [8, 9].

Settee et al., [10] improved the estimate of event importance by traditional event mining and feature generation approaches. Two large real-world news corpora are used for evaluation process. These news corpora have thousands of news article in daily basis. Wikipedia Current Events Portal is used for evaluation. The evaluation process is a large-scale process. The experimental results are compared with language model based ranking techniques and shows the efficiency of the same.

Cataldi et al., [11] detected the most promising topics stated by the community using a traditional topic detection techniques via twitter. It is a real-time process. Traditional aging theory is used to model the term life cycle of the contents extracted from tweets. Term life cycle is used to find the emerging topics. Frequently occurred terms are named as emerging and it was rare in past occurrence. Authority of users are found using Page Rank algorithm. This algorithm is also used to analyze the

social relationship in the network. The contents depends on the sources. Emerging terms are related with semantic keywords using navigable topic graph. This graph is used to find the emerging topics. Validity of the approach is studied in different cases.

Kong et al., [12] differentiated the news topics and user attention by media focus. News influence impact is decided by the media focus via five different strategies. Media focus is combined with user attention by three different strategies. Four different types of interaction between media focus and user attention are considered. This is the first method implemented to compute influence decay of news topics. Ebbinghaus forgetting curve and information fusion are also utilized to assess the performance of the proposed algorithm. Better results were obtained in this method.

Ding et al., [13] predicted the preferences in pairwise comparison by using modeling approach. New generative model is developed for pairwise comparison. This model uses the multiple shared latent rankings. It is also used to predict the behavior of the inconsistent user. The method shows, how the latent rankings in the new generative model is used for predicting the topics. This model tends to be very simple, consistent and provides computational complexity guarantee. Semi-synthetic and real-world datasets is used for performance evaluation. The experimental results shows the improvement in the topic prediction accuracy.

### 3 Proposed Methodology

This section describes about the proposed CFRA based SociRank system. In the sub-sections, the topic identification and ranking of each step has been discussed.

#### 3.1 System Overview

The aim of the proposed method is to identify, consolidate and rank the prevalent topics. It uses both news media and social media. Figure 1 shows the systematic framework of the proposed method. There are four main stages in the proposed architecture.

- (1) *Preprocessing*: In this stage, key terms are extracted from social and news data during particular time. The terms are filtered for further processing.
- (2) *Key Term Graph Construction*: By using the extracted key term, a graph is plotted between key term set and co-occurrence similarity. Key terms are represented in vertices and co-occurrence similarities are represented in the edges. The graph is processed and pruned. After processing, the graph contains cluster of popular topics.
- (3) *Graph Clustering*: Well-defined TCs are found by clustering the graph. HFC algorithm is used to reduce the overlapping in TCs.

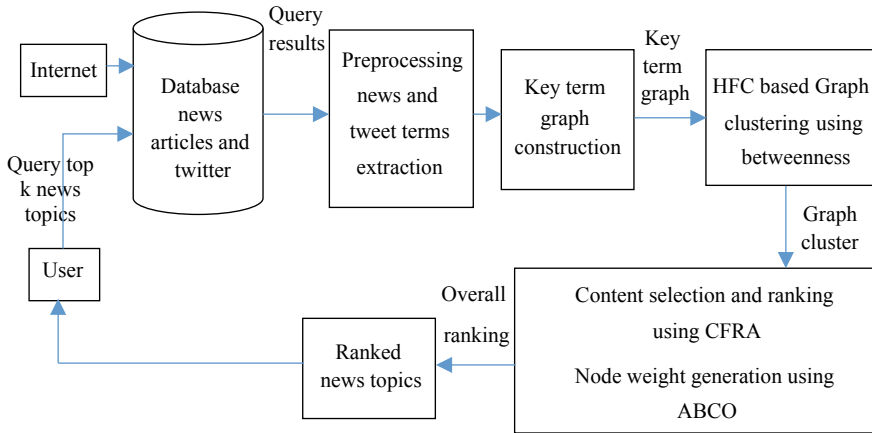


Fig. 1 Overall process of proposed CFRA based SociRank scheme

(4) *Content Selection and Ranking*: Factors like MF, UA and UI are used to select and rank the TCs in the graph. Topic ranking is done by CFRA scheme.

A database is created by collecting the news and tweet data from the internet. Twitter public timeline [14] is used to collect the tweets and RSS feeds of news websites which are used to collect news articles. Top  $k$  ranked news topic in the specified time between date  $d1$  (start) and date  $d2$  (end) are requested by the users.

### 3.2 Pre-processing

The system queries the news articles and tweets in the database. Two of the following terms are created for news article and tweets.

- (1) *News Topic Extraction*: The set of terms from the news data source consists of keywords extracted from all the queried articles.
- (2) *Tweets Term Extraction*: All the unique and relevant terms from the tweet data source are extracted. Identification of tweet language is done at first. Non-English tweets are eliminated. The terms with less than three characters is also eliminated.

### 3.3 Key Term Graph Construction

Clustered nodes in the graph represents the most prevalent news topics. Unique terms in the  $N$  and  $T$  are represented in the vertices and relationship between these terms



are represented in the edges. Method to select the terms and finding the relationship between them is defined in the following section. A graph is pruned after the identification of terms and relationship between them. Unimportant vertices and edges are filtered out.

### 3.4 Graph Clustering Using HFC

After the identification of significant vertices as most significant terms and edges which are used for term-pair co-occurrence values from the graph  $G$ , well-defined TCs sub-graphs, are all identified and separated. The parameters betweenness and transitivity is explained as follows:

- (1) *Betweenness*: Matsuo et al. [15] clustered the co-occurrence graph by an efficient approach. Word clusters are identified by using Newman clustering [16] algorithm. The identification is done by betweenness. It may be expressed as the number of shortest path between pairs of nodes. In, loosely connected clusters by inter-cluster edges, the shortest path between different clusters must go along with these edges. The edges that connects the clusters have high value of edge betweenness. Well-defined clusters are obtained by iteratively removing the edges.

The betweenness measure of an edge  $e$  is defined as [15],

$$betweenness(e) = \sum_{i,j \in V} \frac{\sigma(i, j|e)}{\sigma(i, j)} \quad (1)$$

Set of vertices is represented by  $V$ , number of shortest paths between vertex  $i$  and vertex  $j$  is given by  $\sigma(i, j)$ . The number of those paths that pass through edge  $e$  is given by  $\sigma(i, j|e)$ .

#### Hybrid fuzzy clustering algorithm

In clustering, the set of given objects are assigned into cluster based on the similarity between them. Similarity is identified by the grouping. Same group objects are similar, whereas objects in the different clusters are called dissimilar. Distance function is calculated by the similarity and dissimilarity measure between the objects. Effective information retrieval is done by using this similarity grouping.

Let us say  $C_i, i = 1, 2, \dots, c$  are the cluster prototypes and  $c_i, i = 1, 2, \dots, c$  are the cluster centers. A given algorithm uses a pre-determined distance calculation method (i.e., Euclidean distance) to create  $C_i$  cluster prototypes each having a cluster center,  $c_i$ .

The hard c-means or k-means algorithm is adopted in the proposed method. These algorithm assigns the data point  $x_j$  in a given dataset  $X = (x_1, x_2, \dots, x_n), X \subseteq \mathfrak{R}$  to exactly one cluster. This kind of assigned may be in-efficient if the data point is equal distance from two or more clusters. By assigning this kind of data point to

one cluster and ignoring other cluster is avoided in the fuzzy clustering methods. In fuzzy clustering algorithm, the data point is assigned to more than one cluster with different degree of membership. The membership function is very useful in cluster overlapping.

Different degrees of membership function is based on the fuzzy set theory [17]. In traditional theory, an object either belongs to a set or not. In this, binary form is used to represent the membership function. Let us say  $u_{ij}$  represents the membership of a data point  $x_j$  to a cluster  $C_i$  where  $u_{ij} \in \{0, 1\}$ . This method is followed in hard c-means algorithm. But in Fuzzy set theory, each data point  $x_j$  to a cluster  $C_i$  with a membership degree. The membership vector of the data point is given by,

$$u_j = (u_{1j}, \dots, u_{cj}) \tag{2}$$

Fuzzy partition matrix  $U = (u_{ij}) = (u_1, \dots, u_n)$  is formed for the data set that has  $n$  data points. It is constructed based on the following constraints,

$$\sum_{j=1}^n u_{ij} > 0, \quad \forall_i \in \{1, \dots, c\} \tag{3}$$

$$\sum_{i=1}^c u_{ij} = 1, \quad \forall_i \in \{1, \dots, n\} \tag{4}$$

First constrain states that, all the clusters must have at least one data point. There should not be any empty cluster. Sum of the membership degree of a data to all clusters equals to 1 as stated by the second constraint. These constraints ensure the prevention of full membership of all data points to only one cluster.

Fitness of the semantic cluster is evaluated by the objective function  $J$ . It is used by both hard and fuzzy c-means algorithms. Optimum clusters are obtained by minimizing the cost functions that are interpreted from objective function. Optimal partitioning of given dataset in fuzzy c-means algorithm is done when the following objective function is minimized:

$$J = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||x_j - c_i||^2 \tag{5}$$

Fuzzifier exponent is represented by  $m$  ( $m > 1$ ). Membership degree of the  $j^{\text{th}}$  data point  $x_j$  to the  $i^{\text{th}}$  cluster center  $c_i$  is given by  $u_{ij}$ .

Bezdek et al., [18] optimized the objective function  $J$  by fuzzy c-means algorithm. The optimization is done by updating the fuzzy partition matrix  $U = (u_{ij})$  and cluster center  $c_i$  in each iteration. They are given as,

$$u_{ij} = \frac{1}{\sum_{l=1}^c \left(\frac{d_{lj}^2}{d_{ij}^2}\right)^{1/(m-1)}} \tag{6}$$

$$c_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \tag{7}$$

Fuzzy c-means algorithm is used efficiently in several clustering problems. The performance of it is reduced by two drawbacks. Initial value of the cluster and trapping into local minima reduces the performance. Various implementation of FCM are proposed to improve the performance. QPSO algorithm is proposed here to solve this problem. It has high convergence capability and less parameter when compared to Particle Swarm Optimization algorithm (PSO). QPSO is gradient descent of FCM. It has high global searching capacity and avoids local minima issues.

**QPSO**

PSO is population based optimization method. Birds flocking and fish schooling motivates this algorithm. In this algorithm, swarm represents the population, particle represents the member of a swarm. These are used for optimization. Search direction of particle is found by its previous best particle. Global best particle for all the particle is found. Let  $N$  be the swarm size. Each particle  $i$  ( $1 \leq i \leq N$ ) has two vectors, velocity ( $V$ ) and position ( $X$ ). The position and velocity of each particle in swarm are updated in each iteration as [19],

$$V_{i,j}(t + 1) = w.V_{i,j}(t) + c_1.r_1.(pbest_{i,j} - X_{i,j}(t)) + c_2.r_2.(gbest_j - X_{i,j}(t))$$

$$X_i(t + 1) = X_i(t) + V_i(t + 1) \tag{8}$$

The position of the  $i^{th}$  particle is represented by  $X_i$ . The velocity vectors of the  $i^{th}$  particle is represented by  $V_i$ . Previous best particle of the  $i^{th}$  particle is given by  $pbest_i$ . Global best particle found so far by all particles is given by  $gbest$ . Two independently random numbers with the range of  $[0, 1]$   $r_1$  and  $r_2$  are generated. Inertia weight is given by  $w$ . Acceleration coefficients are given by  $c_1$  and  $c_2$ .

In QBPSO, the particle are considered as sparse data and the fitness evaluation is optimal  $\lambda$  value. A recent theoretical study [20] reported that local attractor is the point convergence of each particle,  $p_i = (p_{i,1}, p_{i,2}, \dots, p_{i,D})$  defined as follows:

$$P_{i,j} = \varphi.pbest_{i,j} + (1 - \varphi).gbest_j \tag{9}$$

where  $\varphi \in (0, 1)$ . Stochastic attractor of particle  $i$  is given by  $p_i$  and it lies in the hyper rectangle with  $pbest_i$  and  $gbest$ .

Sun et al., [21] implemented a quantum-behaved PSO (QPSO) algorithm. In QPSO, each particle only has position vector and does not have the velocity vector. During the evolution, each particle updates its position as follows:

$$X_{i,j}(t + 1) = \begin{cases} P_{i,j}(t) + \beta.(Mbest_j(t) - X_{i,j}(t)).\ln(\frac{1}{u}), & \text{if } h > 0.5 \\ P_{i,j}(t) - \beta.(Mbest_j(t) - X_{i,j}(t)).\ln(\frac{1}{u}), & \text{otherwise} \end{cases} \tag{10}$$

Two random numbers  $h$  and  $u$  in the range of 0 to 1 is distributed uniformly. Contraction-expansion coefficient is given by  $\beta$ . Convergence speed of the algorithm is controlled by this parameter. Mean best position of the population is given by  $Mbest$  and it is calculated as,

$$Mbest_j(t) = \frac{1}{N} \sum_{i=1}^N pbest_{ij}(t) \tag{11}$$

Population size is given by  $N$ . Algorithm 1 defines the main steps in QPSO. Local attractor is given by  $\rho$ . Number of fitness evaluations is given by FEs and maximum number of FEs is represented by MAX\_FEs. Velocity term and the parameters  $w$ ,  $c1$  and  $c2$  are not included in QPSO. But QPSO introduced a new parameter  $\beta$  which is linearly decreased from 1.0 to 0.5 reported in some of the literature [21, 22].

```

Algorithm 1: QPSO for sparse data reduction
Begin
while FEs <= MAX FEs do
for each particle i do
Update the position according to (8);
Calculate the fitness value of the new particle;
FEs++;
end for
Update the pbest, gbest and p in the population;
end while
End
    
```

Development of more efficient clustering algorithm and improvement in the local extreme problem of FCM algorithm are main aim of this research work. A hybrid fuzzy clustering algorithm is proposed for the same.

In this method a set including  $c_i$  cluster centers is represented by a particle. Each particle is constructed by  $x_i = (m_{i1}, m_{i2}, \dots, m_{ic_i})$ , where  $m_{ij}$  represents cluster  $C_{ij}$  with particle's  $j$ th vector of cluster center. Particle's fitness function is given by  $f(x_i) = J_m(U, V) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2(x_i, v_i)$ . The steps involved in this process is explained as follows:

- Step 1: Calculation of initial value of data by the input sample data.
- Step 2: Number of particle swarm (pop-size) confirmation by sample's data
- Step 3: Initialization of the swarm and stochastic. Selection  $c_i$  cluster centers referring to the bound of data for each particle.
- Step 4: The personal value for each particle is initialized and global value of entire swarm is also initialized.
- Step 5: Calculation of the fitness value corresponding to fitness function  $f(x)$  along with current  $pbest$  and global  $gbest$  for each particle is calculated.
- Step 6: The new individual particle  $X_i(t + 1)$  is generated using Eqs. (8), (9), (10);
- Step 7: Go to Step 5 until  $(||f(x_i)^{(j+1)} - f(x_i)^{(k)}||)$  is met, then the algorithm ends.

### 3.5 Content Selection and Ranking Using CFRA Scheme

Prevalent news-TCs are identified for the duration date's  $d1$  and  $d2$ . Contents related to this TCs are selected and ranking is done at last. MF of the topic is represented by the items related to the news media. UA of the topic is represented by the items related o the social media. UA defines the number of unique twitter users. Selection of key term is not an easy process due to the similar key terms that contains irrelevant items.

1) *Node Weighting*: Nodes of each topic is weighted before selecting appropriate content. The terms important to topics are selected by using these weights. These weights represent the topic relevancy. Optimal weight code is generated using ABCO algorithm.

#### ABCO for optimal node weight generation

In ABCO algorithm [23], D-dimensional is mentioned as the solution space of the problem, where D indicates the amount of optimized parameters. ABCO is a selection mechanism for neighborhood of the candidate solutions in the Onlooker Bee (OB) stage.

This selection mechanism was based on data shared by the Employed Bees (EB). By using the EBs average fitness value is calculated and this value are stored into memory. So, the OBs select a neighbor's information from the memory.

The randomly selected site fitness value is given in the following Eq. (12)

$$avg_t^{pop} = \frac{1}{SN} \sum_{i=1}^{SN} fit_i \quad (12)$$

where  $avg_t^{pop} \rightarrow$  an average fitness value of EBs population at iteration  $t$  and  $SN \rightarrow$  number of EBs. EBs fitness values are tested with  $avg_k^{pop}$  and the solutions of EBs, which are better than  $avg_k^{pop}$ , are stored to the board. That board duration of solutions is calculated by

$$D_i = K \cdot fit_i \quad (13)$$

where,  $K \rightarrow$  a positive constant number,  $i \rightarrow$  fitness value of  $i^{th}$  EBs and  $D_i \rightarrow$  waiting time on the memory of the solution and waiting time of the solutions is proportional to fitness values of EBs. Accordingly, neighbors for OBs [ $x_{kj}$  in Eq. (14)] are no longer chosen from the memory.

The amount of food sources are represented by the volume of EB and OB. EB is allocated for each food source. Quantity of the food source equals the overall quantity of the EB. The new sources are obtained using the following equations,

$$v_{ij} = x_{ij} + \varphi_{ij}(x_{ij} - x_{kj}) \tag{14}$$

where:  $i, k = \{1, 2, \dots, SN\}, j = \{1, 2, \dots, D\}$ , randomly generated real-number in the range of  $[-1, 1]$  is represented by  $\varphi$  Randomly chosen index number in the Bee colony is given by  $k$ . Original solution  $v = \{x_{i1}, x_{i2}, \dots, x_{iD}\}$  is compared with the new solution  $v' = \{x'_{i1}, x'_{i2}, \dots, x'_{iD}\}$ . New solution is remembered by the bee if the obtained solution is best suited than the earlier solution. Or else it remembers the former solution. The OB chooses a food source based on the probability and it's provided in Eq. (15):

$$P_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \tag{15}$$

Fitness of the solution  $v$  is given by  $fit_i$ . Number of food sources locations is given by  $SN$ . OB takes the new food source mentioned by Eq. 15. In Scout Bee (SB) phase the food source fitness is not enhanced rather it is discarded. This act signifies the negative feedback in ABCO and that EBs food source happens to be a SB and constructs a random search through Eq. (16)

$$x_{id} = x_d^{\min} + r(x_d^{\max} - x_d^{\min}) \tag{16}$$

Random real number within the range  $[0 1]$  is represented by  $a$ , where  $a = r$ . Lower and upper border in the  $d^{th} \rightarrow$  the problems space of dimension is represented by  $x_d^{\min}$  and  $x_d^{\max}$ . Fitness value with the optimal weights are generated efficiently. The pseudocode of ABCO is given below:

**Algorithm 2: ABCO based optimal weight generation**

Input: Node weight values

Output: Optimal weight selection

1. Initialize the weight parameters values. set the threshold value for SN (Population Size), 50% for employment bees, 50% for non-employment bees, randomly generate the Food Number (SN/2) feasible solutions, the maximum number of iterations is the maxCycle (MCN), the number of stagnation is the limit (during an iteration, if the optimal value does not improved after limit iteration, then reset the feasible solution)
2. the fitness  $fit_i$  of the population is evaluated in glaucoma image
3. for cycle =1
4. Repeat
5. For each EB
  - {
  - New solution  $v_{ij}$  is produced by using equ (14)
  - The fitness  $fit_i$  is calculated
  - Greedy selection process is applied
  - }
6. The probability values  $P_i$  is calculated for the solutions  $i$  by equ (15)
7. For each OB
  - {
  - A solution  $i$  is selected depending on  $P_i$
  - New solution  $v_{ij}$  is produced
  - The fitness  $fit_i$  is calculated
  - Greedy selection process is applied
  - }
8. If an abonded solution occurred for the SB, then, replace it with a new solution which is randomly produced by using equ (16)
9. the best solution is stored until now
10. For cycle=cycle+1
11. Until cycle=MCN.

- (2) *User Attention Estimation*: UA is calculated by collecting the tweets related to the topic and by counting the number of unique user those who have created those tweets. TC of each node is used to ensure the relationship between the tweets and the topics.
- (3) *Media Focus Estimation*: News articles related to TC are selected for the calculation of MF. This presents a problem similar to the selection of tweets when calculating UA. Articles related to the topics are selected by using the weighted nodes of TC. Top  $k$  keywords selected from each article is compared instead of comparing node combinations in the tweet content.  $k$  keywords are selected to get insight into the more important terms of the article. These keyword are used to derive the terms that make up the TCs.
- (4) *User Interaction Estimation*: Degree of interaction between the users those who have created the content of social media is measured. The database is queried for “followed”. These relationship is used to construct the Social network graph. In Twitter, there are two types of relationships: (1) “follow” and (2) followed. In first case, direction of flow of relationship is from  $u1$  to  $u2$ , where as in second case it is from  $u2$  to  $u1$ .

**Collaborative filtering based ranking**

Ranking of the topics are used by the Ranking based RF system to recommend the topics to the users. This method utilize the similarity between two users based on

their ranking. Kendall tau rank correlation coefficient is used as a common measure of similarity.

*Collaborative Filtering*

Set of users is denoted by  $U$ , set of new topics is denoted by  $I$ . For each topic  $u \in U$ , a set of topics  $I_u \subseteq I$  is rated by  $u$  in the recommendation system. Rating matrix is given by  $R$ . For each element  $r_{u,m} \in IN$  is the rating score of the  $m$ th topic  $i_m$  with respect to  $u$ . Different relevance scores are indicated by  $IN$ .

Rating scores predicted by neighborhood users (similar users) are used by the CF to recommend topics. Rating Matrix  $U$  is used to find the similarity between an user  $u$  and each user in  $U$ . Recommendation are given based on the selected neighborhood users  $U_u \subset U$ .

**Ranking-based CF**

Topics are recommended by the Ranking-based CF as per their ranking derived from rating matrix  $R$ . Standard Kendall tau rank correlation coefficient [24] is used to calculate the similarity between two users.

$$\tau_{u,v} = \frac{N_c - N_d}{\frac{1}{2}N(N - 1)} \tag{17}$$

Numbers of the concordant pairs is given by  $N_c$  and number of discordant pairs is given by  $N_d$ .

Indicator function  $sgn_{u,v}(m, n) = 1$  for the concordant topics and it is  $-1$  for the discordant topics.

$$sgn_{u,v}(m, n) = \begin{cases} 1, & \text{if } (r_{u,m} - r_{u,n})(r_{v,m} - r_{v,n}) > 0 \\ -1, & \text{if } (r_{u,m} - r_{u,n})(r_{v,m} - r_{v,n}) < 0 \end{cases} \tag{18}$$

The number of concordant pairs minus the number of discordant pairs ( $N_c - N_d$ ) produces the sum of  $sgn_{u,v}(m, n)$  of all topics.  $\tau_{u,v}$  is given as,

$$\tau_{u,v} = \frac{\sum_{m=1}^N \sum_{n=m+1}^N sgn_{u,v}(m, n)}{\frac{1}{2}N(N - 1)} \tag{19}$$

Preference on a pair of topic of a user  $u$  is calculated by using the preference function as

$$\Psi_u(m, n) = \frac{\sum_{v \in U_u^{m,n}} \tau_{u,v}(r_{v,m} - r_{v,n})}{\sum_{v \in U_u^{m,n}} \tau_{u,v}} \tag{20}$$

Set of similar users of  $u$ , who have rated both  $i_m$  and  $i_n$  is given by  $U_u^{m,n}$ . Preference aggregation algorithm is used to find the total ranking of topics by the predicted pairwise preferences. Users are send with the final ranking results as obtained by the above mentioned procedure.



## 4 Results and Discussion

The performance of the proposed CFRA is evaluated in this section and the performance is compared with SociRank Scheme. Classification context scenario is used to evaluate the performance. Parameters like Accuracy, F1-score, Recall and Precision are used for performance comparison.

### Dataset Description

Tweets from twitter and news articles from news websites from November 1, 2013 and February 28, 2014 are used as a datasheet. In this period, 175,044,074 bilingual tweets and 105,856 news articles are collected. From this only 71,731,730 English tweets are used. Other tweets are discarded due to the non-English language. Websites like cbsnews.com, andusatoday.com and bbc.com are used for article collection. This dataset is divided into two categories.

- (1) Data from January and February 2014 is used as a testing dataset. Overall method is evaluated using this dataset.
- (2) Data from November and December 2013 is used as a control dataset. The thresholds for producing best results are established by performing experiments on this dataset.

### Performance Evaluation

Figure 2 represents the percentage of topic selection with overlap voted topics by MF, SociRank and CFRA based SociRank Method. It indicate that, the CFRA based SociRank produces good results. Voted topics indicates the topics selected by the users as an important one. CFRA based SociRank is best suited for news topic discovery and it produces very good results when compared with the techniques that uses only the news media data.

Figure 3 represents the overlap percentage between top 10 topics and top 10 voted topics by MF, SociRank and CFRA based SociRank Method. Green line in the

**Fig. 2** Comparison of overlap percentage

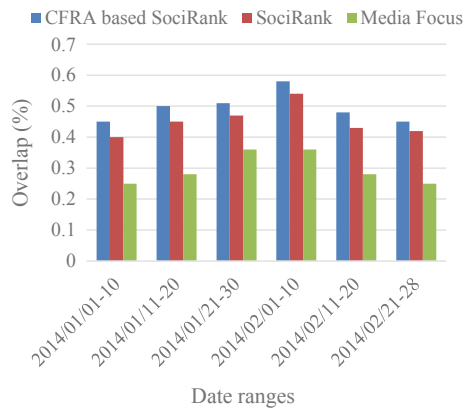
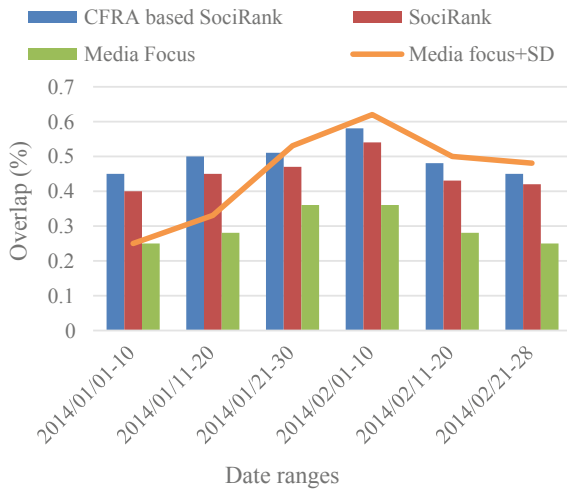


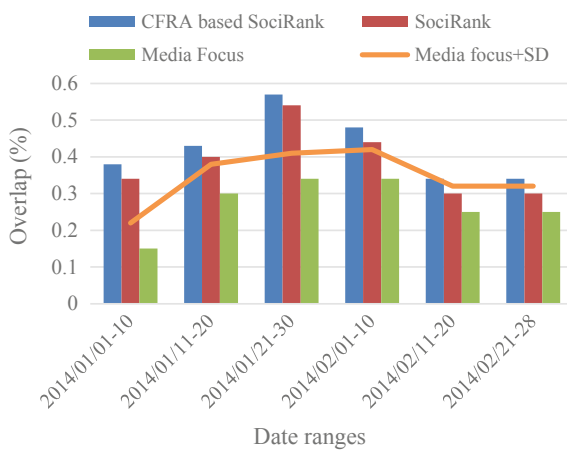
figure indicates the overlap percentage of MF with 1 standard deviation. Significant improvement of overlap percentage is considered if the SociRank bar surpasses the green line. But, in this figure it does not occur. So there is no significant difference between the approaches.

Figures 4, 5 and 6 indicates the overlap percentage of top 20, 30 and 40 ranked list. In this also CFRA based SociRank have produced good results. The experimental results of CFRA based SociRank is high when compared to the SociRank and MF methods. In Fig. 7 average overlap percentage between elected top K and selected top K between 10 to 40 ranks list are listed.

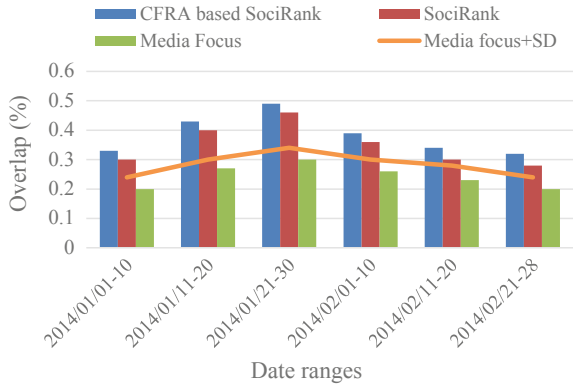
**Fig. 3** Overlap percentage between top 10 voted topics and top 10 topics



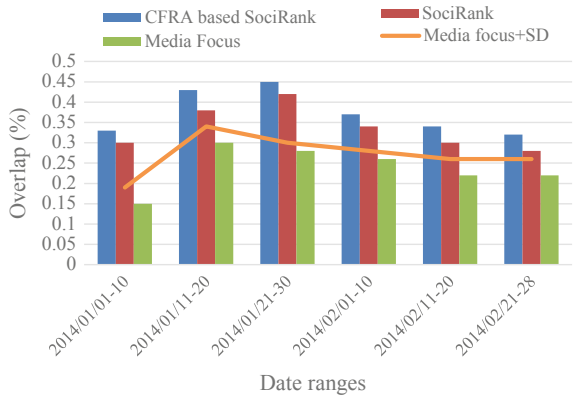
**Fig. 4** Overlap percentage between top 20 voted topics and top 20 topics



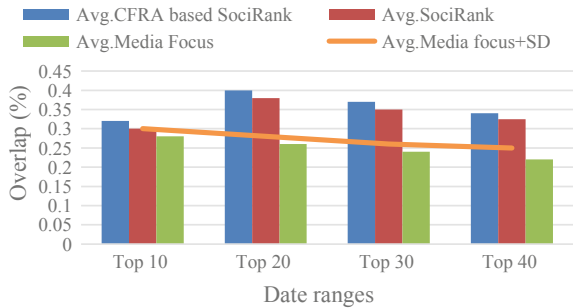
**Fig. 5** Overlap percentage between top 30 voted topics and top 30 topics



**Fig. 6** Overlap percentage between top 40 voted topics and top 40 topics



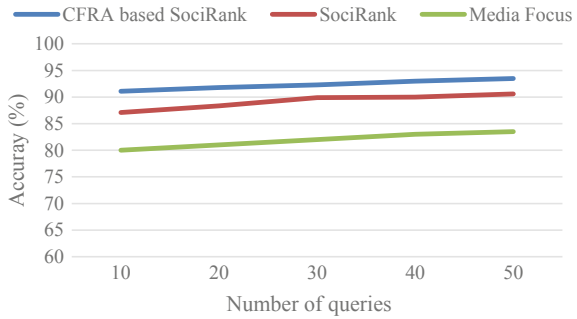
**Fig. 7** Average overlap percentage between elected top K and selected top K



Average overlap percentage of top 10, 20, 30 and 40 ranked list is represented in Fig. 7. In top 20, 30 and 40 ranked list, the CFRA based SociRank method significantly outperforms the SociRank and MF method.

The accuracy of all topic ranking schemes graphical representation is shown in Fig. 8. It shows the accuracy of proposed scheme CFRA based SociRank is 2.9%

**Fig. 8** Performance comparison of accuracy among all topic ranking schemes

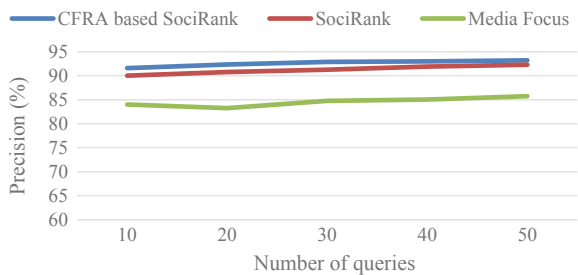


and 10% expanded than SociRank and MF. The proposed CFRA based SociRank attained high accuracy rate of 91% compared with other schemes, due to the effectual graph clustering and optimal weight generation.

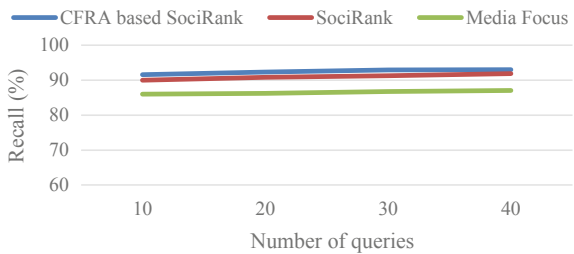
The precision of all topic ranking schemes graphical representation is shown in Fig. 9. It shows the precision of proposed scheme CFRA based SociRank is 0.9% which is 7.44% higher than existing SociRank and MF schemes. Due to high positive rate, the precision of proposed scheme attained a high value.

The recall of all topic ranking schemes graphical representation is shown in Fig. 10. It shows the recall of proposed scheme CFRA based SociRank is 1.07% which is 5.78% higher than existing SociRank and MF schemes. Because of the less

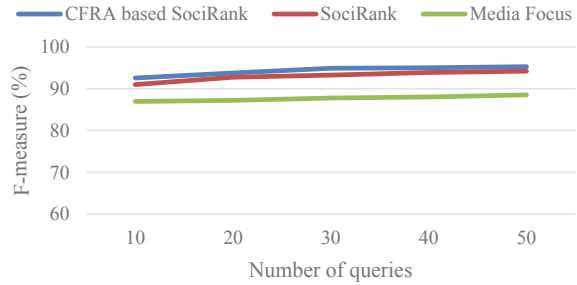
**Fig. 9** Performance comparison of precision among all topic ranking schemes



**Fig. 10** Performance comparison of recall among all topic ranking schemes



**Fig. 11** Performance comparison of F-measure among all topic ranking schemes



error rate and high precision esteems, the review of proposed CFRA based SociRank is better than other techniques.

The F-measure of all topics ranking schemes graphical representation is shown in Fig. 11. It shows the F-measure of proposed scheme CFRA based SociRank is 1.1% and it is 6.74% higher than existing SociRank and MF schemes. Due to the high precision and recall rate, the F-measure of proposed attained high and the effectual clustering improved the accuracy result of this scheme.

## 5 Conclusion

In this paper, a new online news topic ranking algorithm has been proposed for finding top K news headlines on real-time social media news topics. Here the focus is based on inconsistency analysis between Media Focus (MF) and User Attention (UA). Overlapping Topic Clusters (TCs) are obtained using Hybrid Fuzzy Clustering (HFC) approach. Secondly, node weight generation using the Artificial Bee Colony Optimization (ABCO) is performed. Individual users are presented with topics differently by personalizing SociRank. Here in this case the ranking is performed using the Collaborative Filtering based Ranking Algorithm (CFRA). News topics are identified automatically and the quality is improved by CFRA based SociRank. Topic trends in continuous time slot can be analyzed in the near future. User behavior can also be used in future for top ranking.

## References

1. <http://www.nist.gov/speech/tests/tdt/>
2. C. Wang, M. Zhang, S. Ma, L. Ru, Automatic online news issue construction in web environment, in *Proceedings of the 17<sup>th</sup> International Conference on World Wide Web*, pp. 457–466 (2008)
3. J. Kleinberg, Authoritative sources in a hyperlinked environment, in *Proceedings of 9<sup>th</sup> ACM-SIAM Symposium on Discrete Algorithms* (1998)
4. G.P.C. Fung, J.X. Yu, P.S. Yu, H. Liu, Parameter free bursty events detection in text streams, in *Proceedings of the 31<sup>st</sup> VLDB Conference*, pp. 181–192, Trondheim, Norway (2005)
5. K.Y. Chen, L. Luesukprasert, S.T. Chou, Hot topic extraction based on timeline analysis and multi-dimensional sentence modeling. *IEEE Trans. Knowl. Data Eng.* **19**(8), 1016–1025 (2007)
6. T. He, G. Qu, S. Li, et al., Semi-automatic hot event detection, in *Proceedings of the 2<sup>nd</sup> International Conference on Advanced Data Mining and Applications*, LNAI4093, pp. 1008–1016 (2006)
7. D. Blei, Probabilistic topic models. *Commun. ACM* **55**(4), 77–84 (2012)
8. S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, M. Zhu, A practical algorithm for topic modeling with provable guarantees, in *Proceedings of the 30<sup>th</sup> International Conference on Machine Learning*, Atlanta, GA, USA, (2013, June)
9. W. Ding, M. H. Rohban, P. Ishwar, V. Saligrama, Efficient distributed topic modeling with provable guarantees, in *Proceedings of the 17<sup>th</sup> International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, (2014, April)
10. V. Setty, A. Anand, A. Mishra, A. Anand, Modeling event importance for ranking daily news events, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 231–240, ACM (2017, February)
11. M. Cataldi, L. Di Caro, C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, in *Proceedings of the tenth international workshop on multimedia data mining*, p. 4, ACM (2010, July)
12. L. Kong, S. Jiang, R. Yan, S. Xu, Y. Zhang, Ranking news events by influence decay and information fusion for media and users, in *Proceedings of the 21<sup>st</sup> ACM international conference on Information and knowledge management*, pp. 1849–1853, ACM (2012, October)
13. W. Ding, P. Ishwar, V. Saligrama, A topic modeling approach to ranking, in *Artificial Intelligence and Statistics*, pp. 214–222 (2015, February)
14. Twitter. [Online]. Available: <http://www.twitter.com>. Accessed Feb 2014
15. Y. Matsuo, T. Sakaki, K. Uchiyama, M. Ishizuka, Graph-based word clustering using a Web search engine, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 542–550 (2006)
16. M. Girvan, M.E.J. Newman, Community structure in social and biological networks, in *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826 (2002)
17. H.-H. Chen, M.-S. Lin, Y.-C. Wei, Novel association measures using web search with double checking, in *Proceedings of the 21<sup>st</sup> International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 1009–1016 (2006)
18. J.C. Bezdek, R. Ehrlich, W. Full, FCM: the fuzzy c-means clustering algorithm. *Comput Geosci* **10**(2–3), 191–203 (1984)
19. Y. Shi, R. Eberhart, Modified particle swarm optimizer, in *Proceedings of the IEEE International Conference on Evolutionary Computation (ICEC '98)*, pp. 69–73, (1998, May)
20. F. Vandenbergh, A. Engelbrecht, A study of particle swarm optimization particle trajectories. *Inform Sci* **176**(8), 937–971 (2006)
21. J. Sun, B. Feng, W.B. Xu, Particle swarm optimization with particles having quantum behavior, in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '04)*, pp. 325–331, (2004, June)

22. J. Sun, W.B. Xu, W. Fang, A diversity-guided quantum behaved particle swarm optimization algorithm, in *Simulated Evolution and Learning*, vol. 4247 of Lecture Notes in Computer Science, pp. 497–504, Springer, New York, NY, USA (2006)
23. D. Karaboga, B. Akay, A comparative study of artificial bee colony algorithm. *Appl Math Comput* **214**(1), 108–132 (2009)
24. J.I. Marden, *Analyzing and Modeling Rank Data* (Chapman & Hall, New York 1995)

# Diversity Among Finger Millet Accessions Based on Genotyping Potential of SSR, EST-SSR and ISSR Markers



Bheema Lingeswara Reddy Inja Naga and S. Sivaramakrishnan

**Abstract** Genetic diversity is the basis for adaptability of any organism against environmental changes through natural selection. Microsatellites markers are mostly used for studying population genetic diversity because of their locus-specific and hyper-variability. Availability of sequences (DNA) paved a way for the development of EST-SSRs. The ISSR markers were identified based on the amplification regions; usually 100–3000 bp present between inversely oriented closely spaced microsatellites. 40 SSR markers, 33 EST-SSR markers, 10 ISSR markers were used in the current study to check diversity among 30 diverse finger millet accessions from different geographical regions. 29 SSR markers, 21 EST-SSR markers and 10 ISSR markers exhibited polymorphism in which 19 SSR markers, 13 EST SSR markers and 8 ISSR markers have a PIC value above 0.5. 30 accessions are categorized into four major groups basing on the distribution of their polymorphic alleles. This current study demonstrates the use of different primers in studying diverse finger millet accessions. These identified primers will be valuable sources for any further genetic studies on Finger millet.

**Keywords** Finger millet · SSR · EST SSR · ISSR · Genetic diversity

## 1 Introduction

Finger millet known as *Eleusine coracana* (L.) Gaertn (Scientific Name) is a very important cereal crop in Sri Lanka, India and Eastern Africa. It has some outstanding properties as an subsistence food crop; especially seeds are rich in calcium, methionine and iron [1, 2] as well as  $\alpha$ -linolenic acid which is essential for human beings [3].

---

B. L. R. I. Naga (✉)  
SSIIE-TBI, SPMVV, Tirupati, India  
e-mail: [reddyinja@gmail.com](mailto:reddyinja@gmail.com)

S. Sivaramakrishnan  
RGCB, Poojappura, Thiruvananthapuram, Kerala 695014, India  
e-mail: [siva\\_ram50@hotmail.com](mailto:siva_ram50@hotmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_27](https://doi.org/10.1007/978-3-030-46939-9_27)



DNA markers have accelerated the pace of genetic analysis/studies by their enormous applications. SSRs (simple sequence repeats) are a class of repetitive DNA sequences which are expressed differently within populations and among different species. There are currently very few informative SSR markers available to study finger millet when compared to crops like rice [4–8]. Panwar Panwar et al. [9] used RAPDs and single sequence SSRs and Cytochrome P450 gene markers for studying diversity among various finger millet accessions. Dida et al. [4] used 45 SSR based markers to study seventy-nine finger millet accessions from Asia and Africa. Upadhyaya et al. [10] at ICRISAT-INDIA used 20 SSR based markers for studying over Nine Hundred and Fifty-Nine finger millet accessions.

ESTs available in the public database are mined using various computational approaches leading to a faster and more economical markers development. ESTs in particular are used for markers development as they represent the coding regions of the genome [11] and also the availability of microsatellites is very high in ESTs [12, 13]. Obidiegwu et al. [14] used three selected EST-SSR markers to characterize forty-eight accessions from the finger millet core set. Reddy et al. [15] used 30 EST-SSR markers to study diversity among 15 finger millet accessions.

Inter simple sequence repeats (ISSRs) is an microsatellite-based multi locus marker technique, which are useful for determining genetic diversity in various crops [16]. Microsatellites are dispersed throughout the genome, whereas the regions with higher abundance are named as “SSR hot spots” [17] which are sources of ISSRs. Few studies have used ISSR to study diversity among finger millet accessions [18, 19].

The objective of the current study is to validate the polymorphic potential of three different marker types (SSRs, EST-SSRs and ISSRs) by studying the diversity among 30 selected finger millet accessions.

## 2 Materials and Methods

### SSRs

SSR markers were developed using the streptavidin coated magnetic bead capture method [7].

### EST-SSRs

Sequences from EST NCBI database are used for developing the SSR markers. “SSRIT” (SSR repeat finder) was used for SSR identification and primer designing was done using Primer3 software [15].

### ISSRs

10 ISSRs were designed based on the repeats which were mostly found in SSR sequence.

### Plant material

30 different accessions of finger millet which were very diverse phenotypically were collected from RARS, Vizayanagaram. Phenotypic data collected in the field.

### Diversity analysis using molecular markers

DNA were isolated using CTAB method with few modifications [20]. PCR analysis was done using three type of markers—40 SSRs (20 SSR primers were developed from the current study and remaining 20 from different loci of finger millet map [4], 10 ISSRs and 33 EST-SSRs [15]. The 20  $\mu$ l PCR reaction consisted of buffer (1X), dNTPs each (0.20 mM), 5 pmol primers, template DNA (20 ng), Taq DNA polymerase (0.5 units). PCR was done in Thermocycler (MJ Research, USA) machine; Initial denaturation of 95 °C for 3 min, next 36 cycles of annealing at 95 °C for 45 s, 59 °C for 45 s (different annealing temperatures used based on primer), 72 °C for 1.5 min and a final extension of 72 °C for 6 min. Gel electrophoresis carried out at 100 V for 3 h. later Gels were visualized and the amplified DNA fragments were scored.

### Scoring and data analysis

Bands which are reproducible were scored. PCR was repeated and the bands were scored by 2 different individuals to ensure reproducibility of the results. Bands are scored based on presence indicate as 1 and absence indicated as 0 at each band position for all the 30 samples and for each primer. 100 bp ladder (NEB, USA) was used as standard. The PIC value was estimated for each primer [21]. PIC is estimated using  $PIC = \Sigma(1 - p_i^2)/n$ , where n is the number of band positions analyzed and  $p_i$  is the frequency of the banding pattern.

### Similarity matrix and cluster analysis

NTSYS-pc 2.0 [22] was used for the analysis. Similarity matrices were constructed from all primers and the matrix of similarity coefficients were subjected to UPGMA to generate a dendrogram. To compare diversity within and among accessions, similarity matrix was subjected to PCA (principal coordinate analysis) [23]. The clustering was done using Unweighted Pair Group Method [24].

## 3 Results

The partial genomic library consisted of ~500 clones with an average size of 350 bp out of which 80% were discarded as false positives after PCR and Dot blot hybridization. Clones with  $\geq 10$  repeats and sufficient flanking sequence for designing primer were used to design 20 primers which were used for diversity studies while the rest were discarded (Table 1). 132 primers were developed from sequences having repeats of  $\geq 5$  and also that the repeat length was above 15 bases [15]. 33 primers are used for the current study diversity study based on their high repeat number (Table 2).

**Table 1** List of SSR primers used in the study

Primer name	Forward primer	Reverse primer	Size	PIC
FMSR1	AAGATCGAAACAAGCAAAAACA	GAAAGAGTATGTGTTGCTTG	100	0.7
FMSR2	TGGAAACAAGCAAAAGATAC	GTATGTTTTGCTTGTTCGA	100	MO
FMSR 3	AAGATCGAAACAAGCAAAAAG	GAGTATCTTTTGTCTGTTC	100	MO
FMSR 4	GATGGGAGACAAGCCAA	ACCTTTTGTCTGTATGGATC	100	MO
FMSR 5	AAGATCCATACAAGCAAAAAG	TCTTTTGTCTGTTTCGATCT	120	0.71
FMSR 6	AGATGTGACACCGAAACTAG	GGACCAAAAATACAGACAAG	190	0.57
FMSR 7	GGAATGGATGTGGTGGTG	TCCTTATAAAAGGAGACATGG	120	0.66
FMSR 8	TCATTCCGAGTTGTCACCAA	CTCAAGCTATGCATCCAACG	100	MO
FMSR 9	CGGCCCGAATACTACTA	ATGTGCGTCAGACTCAATGG	150	0.6
FMSR10	TAGCAGCAGTAGCAGCAGCA	CGAGCGTACTGTGTTGTGT	100	0.5
FMSR11	TCCGCCAATTCTATTCTGTC	GTGCGTGGAGCGATTATTCT	100	0.3
FMSR12	CGGCCCGAATTCTACTA	GTTTTCCAGTCACGACGTT	100	0.31
FMSR13	GCGAATACACTAGTAGATATAGCAGCA	ATGTGCGTCAGACTCAATGG	100	0.19
FMSR14	TGGATGTCGATGCTGTTTGT	CACACGTCCAAGGGAGTTT	100	0.52
FMSR15	CGGTGCTGTGTTGTTTCTG	AGGGAAGAAAGCGAAAGGAC	120	0.3
FMSR16	TTAAGAACCACCGCAAAAC	TGTGGAATTGTGAGCGGATA	150	MO
FMSR17	GGGAAGTCTTGTGTGTCAT	ATCCTCTACACCGCTTTT	150	MO
FMSR18	GCTAGCAGAATCACTCCGTGT	TTGCAAATCTTCACCGCTAT	100	0.54
FMSR19	GACGCTCTGCAAAATCAGT	TGTGGAATTGTGAGCGGATA	100	0.67
FMSR20	AGCTAGCAGAATCCGACTGG	TGTGGAATTGTGAGCGGATA	100	0.5
UGEPI	TTCAGTGGTGACGGAAGTTCT	GGCTCCATGAAGAGCTTGAC	200	0.6392
UGEPI5	TGTACACAACACCACACTGATG	TTGTTTGGACGTTGGATGTG	250	0.67
UGEPI6	AGCTGCAGTTTCAGTGGATTC	TCAACAAGGTGAAGCAGAGC	220	0.55
UGEPI8	ATTTCCGCCATCACTCCAC	AGACGCAAATGGTAAATGTC	300	0.28
UGEPI12	ATCCCCACCTACGAGATGC	TCAAAGTGATGCGTCAGGTC	210	0.8767
UGEPI21	CAATTGATGTCATTGGGACAAC	GTATCCACCTGCATGCCAAC	200	MO
UGEPI26	ATGGGGTTAGGGTTCGAGTC	TGTCCCTCACTCGTCTCCTC	200	0.712
UGEPI31	ATGTTGATAGCCGAAATGG	CCGTGAGCCTCGAGTTTATAG	300	0.7057
UGEPI53	TGCCACAACCTGTCAACAAAAG	CCTCGATGGCCATTATCAAG	200	0.7441
UGEPI56	CTCCGATACAGGCGTAAAGG	ACCATAATAGGGCCGCTTG	250	MO
UGEPI65	AGTGCTAGCTTCCATCAGC	ACCGAAACCCTTGTGAGTTC	200	0.4754
UGEPI68	CGGTCAGCATATAACGAATGG	TCATTGATGAATCCGACGTG	250	MO
UGEPI77	TTCGCGGAAATATAGGC	CTCGTAAGCACCCACCTTTC	200	0.651
UGEPI78	AAGCAATCAACAAAGCCTTTTC	TACAACGTCCAGGCAACAAG	200	0.4
UGEPI81	AAGGGCCATACCAACTCC	CACTCGAGAACCGACCTTTG	200	0.31
UGEPI102	ATGCAGCCTTGTATCTCC	GATGCCTTCCTTCCCTTCTC	200	MO
UGEPI106	AATCCATTCTCTCGCATCG	TGCTGTGCTCCTCTGTTGAC	200	0.13
UGEPI107	TCATGCTCCATGAAGAGTGTG	TGTCAAAAACCGATCCAAG	200	0.1
UGEPI108	GTTGGCTGCTGCTTATCC	TATCTGCTTGTGCAGCTTCG	200	MO
UGEPI110	AAATTGCGATCCTTGCTGAC	TGACAAGAGCACACCGACTC	200	0.63

**Table 2** List of EST-SSRs used in the present study

Primer	Sequence	Repeat	Size	PIC
FMESTSSR1	GAAGTGTGGGGAGTGAAAT	(GA)16	450	0.9
	CCGCACATTACCCTCTCATT			
FMESTSSR2	TGGAAAAGGGAAAATCGTGA	(GA)11	200	MO
	ATCTCATCTCGCCACCACTT			
FMESTSSR3	GGAAAAAGGGAAGGATCAGG	(GA)10	190	MO
	CTCACTCCTCTTCGCTGACC			
FMESTSSR4	AGCGAGATGTCCGTGAGACT	(GA)10	150	MO
	CGCTCTCGCTCTTACTCTCG			
FMESTSSR5	GGGGATAGAAAGAGAGG	(GA)12	200	MO
	GATCCAATCAGCCGCACTAT			
FMESTSSR6	TAACCTGGGAAGAGCGAGAA	(TCT)8	200	MO
	CCCCCAACAATAATGCCTAA			
FMESTSSR7	TGAGGCCTCTCCATATCCAC	(AGC)7	200	MO
	CAGGCCGAGAGAAAGAGAGA			
FMESTSSR8	ATCGAGGCGATGAGAGTTTG	(TTC)8	200	0.2
	AAATGGCCAAACGAAACAAC			
FMESTSSR9	GGGACTCTAGTTCCGCTTTC	(TC)15	200	0.7
	AAGCTCAAATCCACACGTC			
FMESTSSR10	GGCGGCTGCTAGGGTTC	(GAGGC)5	300	0.85
	CGCTCAATCATGACAACAC			
FMESTSSR11	TCCCTCCTCTCATCCTCTGA	(CT)17	400	0.22
	GGCAAATTCGATTGAGGCTA			
FMESTSSR12	GCTGAGTCGTACCGAGATTAGTT	(GA)12	400	0.22
	CGACGACGAGTCGTACTIONTGA			
FMESTSSR13	GCCACTCGAAACGCAAG	(CCT)7	300	0.43
	GAAACGGTGCAGCCTCTTAG			
FMESTSSR14	AGATCGGCAGCCACTACATC	(CGT)7	400	0.8
	GAGACTGAGAAGCCGTGCAT			
FMESTSSR15	CGTCGATCAGTCAGTCATGC	(TCCC)5	400	0.8
	CATGGGGTTGATCTTGAGAGA			
FMESTSSR16	GAGGCATGCACGTACAACAC	(GCG)7	400	0.8
	GGAGGGAGGGAATTCACAAT			
FMESTSSR17	CATCTCCATCTCCATCTCCA	(AAGAG)7	400	0.3
	AAGGACGATCGCAACCAG			
FMESTSSR18	CATCTCCATCTCCATCTCCA	(AAGAG)7	200	0.5
	GGACTTGAGGCAGTTGCAG			

(continued)

**Table 2** (continued)

Primer	Sequence	Repeat	Size	PIC
FMESTSSR19	AGATCGGCAGCCACTACATC	(CGT)7	200	MO
	ACTGAACCAAGATCCGATGC			
FMESTSSR20	TCTCCATCTCCATCTCTACTCG	(AAGAG)7	200	MO
	GGACTTGAGGCAGTTGCAG			
FMESTSSR21	CGTAGTAGTACATCACAGCTA	(AAGAG)7	200	0.9
	CTGATGGCGTATGGGAGTCT			
FMESTSSR22	CACTACACCGCATCATCTCG	(AGA)18	200	0.6
	AGCCGTGATGCCTACAACCTC			
FMESTSSR23	GCGAGTGAGAGAGGGAGCTT	(AG)16	200	0.81
	GTCCAGCTGTTGCTGTTGAA			
FMESTSSR24	ATGGACCAAGAAACCTCACG	(GCG)7	300	0.25
	TCCTCGAACGGGAATCTCTA			
FMESTSSR25	TCCATCATCCATCTCCATCTC	(AAGAG)7	400	0.83
	GGCGTTGAGGCTCCTGAC			
FMESTSSR26	CATCTCCATCTCCATCTCCA	(AAGAG)7	150	0.4
	GGACTTGAGGCAGTTGCAG			
FMESTSSR27	AGAAGGCCCCCGATTTATTT	(GA)10	200	MO
	GGGTTGTGGCTGTTGGTAGT			
FMESTSSR28	AGGAGCCTAGGACGAACTCC	(GAG)7	200	0.6
	CCTCCTCCTCCTCCTCATCT			
FMESTSSR29	CCACCTGCTCCATCTACATCT	(AAGAG)7	200	0.7
	AAGGACGATCGCAACCAG			
FMESTSSR30	AAGGGTCTGCTGCTGTGAGT	(CCA)7	100	MO
	TGGTGTGTTGTCTCGGTGGTA			
FMESTSSR31	CAGGCGGCTAAGGTAGTGAG	(GGC)6	200	0.9
	GATGTAGTCCGGCAGGTAGC			
FMESTSSR32	GATCGTTTCCGATCGAGTGT	(TC)9	200	MO
	CGCCCAGAGGTAGCATATAAA			
FMESTSSR33	TCATTGCATGGGATGAAGAA	(GT)9	300	MO
	CAGGCCAGCATTACACACAC			

Based on the EST-SSR and SSR data it is evident that CT and GA were most found repeats in finger millet based on which 10 ISSR primers were designed (Table 3).

Phenotypic data (Table 4) reveals that accession 3721 has the highest days to flowering and accession 5817 had the least, only 6 accessions (3391, 4565, 7018, 3721, 5201 and 2790) were found to be pigmented; 2457 is the tallest while 6221 was the shortest; Highest value was found in following accessions: FLBL: 3721, EXER: 3721, LLF: 2790, FLBW: 3721, peduncle length: 2871, PBN: 5870, plant

**Table 3** List of ISSRs used in the present study

Primer	Sequence	Size	PIC
FMISSR1	(CT)6TTG	300–1500	0.6
FMISSR2	(CT)6TA	350–1300	0.8
FMISSR3	(CT)6TG	1000–1200	0.4
FMISSR4	(CA)6AC	400–1000	0.7
FMISSR5	(CA)6AG(GT)	200–1000	0.65
FMISSR6	(CA)6T	700–1300	0.35
FMISSR7	(CAC)3CC	220–1000	0.8
FMISSR8	(GA)6GG	120–1200	0.7
FMISSR9	(GA)6CC	220–1000	0.75
FMISSR10	(GA)6T	150–1000	0.8

yield: 3475, FLSL: 6165, inflorescence length: 2790, inflorescence width: 501, WLF: 2034, PAS: 501 accession.

29 SSRs primers showed polymorphism out of the 40 SSR primers tested. 14 primers of 20 primers designed in current study exhibited polymorphism (10 primers—PIC value of  $\geq 0.5$ ); 15 primers taken from [4] exhibited polymorphism (9 primers—PIC value of  $\geq 0.5$ ). overall 19 SSR markers had PIC value of  $\geq 0.5$  out of 29 SSR markers tested. Most SSR markers amplified in the range of 100–200 bp (Table 1).

21 primers showed polymorphism out of the 33 EST-SSRs used in the study. Most primers showed amplified products in range of 100–450 bp size (average size of 200 bp). 14 primers of the 21 polymorphic primers exhibited PIC values of  $\geq 0.5$  and 9 had PIC value of 0.8. The primers FMESTSSR 1, 20, 31 exhibited highest PIC (0.9) (Table 2).

All 10 ISSRs showed polymorphism out of which 8 exhibited PIC value of  $> 0.5$ . Primers amplified fragments ranging from 200 to 1200 bp (Table 3). Dendrogram was constructed based on scoring revealed four groups (Fig. 1). Accessions 2871 and 2457 grouped as one. The second group had 3 sub-groups in which the first one had 3392, 6165, 3317, 4497 and 5870 accessions, second had 4757, 3165, 4734, 6154, 6082, 4565, 6337, 3391 accessions in which 4757 and 4734 accessions (both from India) were similar, and the third has 2 accessions (6350, 7018); most of the African accessions got clubbed under this group. The third group has 3 subgroups in which first had 14, 121, 3614, 4491, 5817, 5201, 2093, 3470, 3721, 501 and 6221 accessions, second had 2790 accession and the third had 3475 accession; Most Indian accessions got clubbed under this group. The 4th group had one accession from India (2034). The PCoA plot is similar to the dendrogram from NTSyS-pc 2.0 (Fig. 2). Accession 2034 got separated from the other accessions. EST-SSR and ISSR were found to be most polymorphic (Fig. 3).

**Table 4** Phenotypic data of the 30 accessions used in the study

	Core	DFP	PP	PLHT	GH	BT	CB	FLBL	FLBW	FLSL	Pend
1	2457	66	g	130.0	e	2.0		35.0	1.3	9	20.0
2	2871	69	g	105.0	e	5.0	y	39.0	1.0	8.5	27.0
3	3392	66	g	74.0	e	3.0		35.0	1.0	9.0	19.0
4	4757	60	g	87.0	e	6.0	y	36.0	1.2	9.0	22.0
5	3317	70	g	104	e	8.0	y	22.0	1.3	8.0	18.5
6	3165	70	g	113.0	e	4.0		22.0	1.0	9.0	20.0
7	4734	43	g	93.0	e	6.0		24.0	0.8	11.0	24.0
8	6165	57	g	91	e	4	y	35	1.6	12	23
9	3391	72	p	114	e	3.0		35.0	2.0	11.0	24.0
10	4497	72	g	117	e	2		32.0	1.5	9.0	22.0
11	5870	57	g	87	e	4		30	1	9	15
12	6082	53	g	94	e	6	y	29	1.3	9.5	13
13	4565	68	p	99.0	e	4.0	y	32.0	1.0	9	20.0
14	6154	62	g	90.0	e	3.0	y	42.0	1.4	8.0	18.0
15	6337	62	g	90.0	e	3.0	y	34.0	1.3	9	26.5
16	6350	70	g	74.0	e	3.0	y	36	1.0	8	12.0
17	7018	59	p	105.0	e	5.0		36.0	1.0	7	23
18	4121	65	g	79	e	6.0	y	35.0	1.5	8.0	9.0
19	4491	63	g	115	e	3		35	1	10	26
20	3614	65	g	138	e	1.0		40.0	0.9	9.0	17.0
21	2034	59	g	97.0	e	10.0		35.0	1.0	8.0	20.0

(continued)

Table 4 (continued)

	Core	DFP	PP	PLHT	GH	BT	CB	FLBL	FLBW	FLSL	Pend
22	3470	59	g	90.0	e	5.0		37.0	0.9	11.0	23.0
23	6221	55	g	55.0	e	5.0		35.0	1.0	9.0	13.0
24	3721	82	p	132	e	3.0	y	48.0	2.2	10.0	30.0
25	501	41	g	103	e	6.0	y	45.0	2.0	7.0	19.0
26	5201	62	p	120.0	e	13.0	y	36.0	1.0	15	27.0
27	5817	40	g	88	e	6.0	y	32.0	1.5	11.0	21.0
28	2093	46	g	73.0	e	9.0	y	34.0	1.1	9.0	17.0
29	3475	59	g	85.0	e	10.0	y	38.0	1.0	10	27.0
30	2790	65	p	104.0	e	3.0	y	33.0	1.0	10	20.0
	EXER	INFLL	INFLW	LLF	WLF	PBN	Plyd	Infic	LO	PAS	GRC
1	11	9	4	8.5	1.1	7	28	Mo	B	3	Rb
2	18.5	5	3	4.0	1.0	9	38	Fis	A	2	Lb
3	10.0	6.5	3.0	3.0	1	7	24	tc	A	3	lb
4	13.0	4.5	4.0	4.5	1	7	34	tc	A	4	db
5	10.5	6.5	3.5	6.5	1	7.0	20	tc	A	4	lb
6	11.0	8.5	4.0	4.5	0.9	7	24	tc	A	2	lb
7	13.0	4.5	4.0	4.5	0.5	9	20	I	A	3	rb
8	11	5.5	6	5	0.9	12	10	tc	A	4	db
9	13.0	6.5	4.5	5.0	0.8	7.0	22	tc	A	3	lb
10	13.0	9.5	4.0	4.0	0.8	9	34	tc	A	3	lb
11	6	4.5	3.5	3	0.9	16	14	tc	A	2	lb

(continued)



Table 4 (continued)

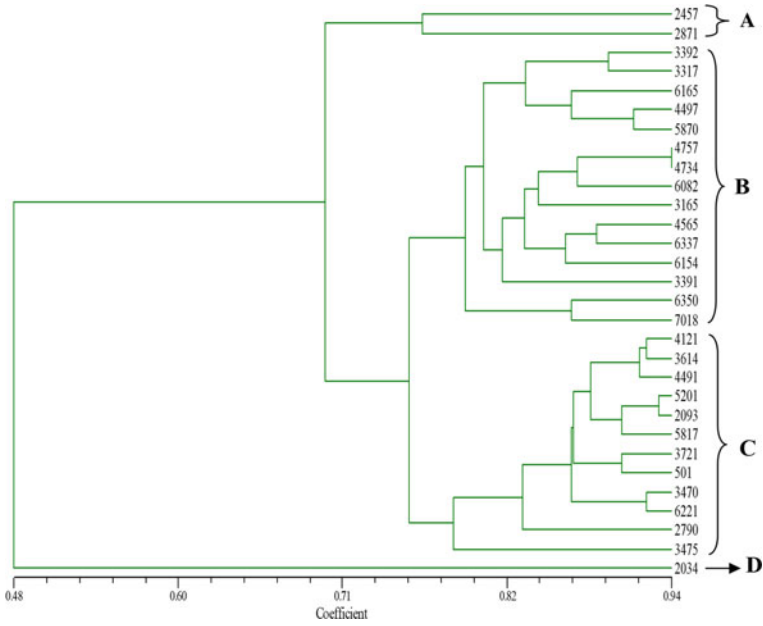
	EXER	INFLL	INFLW	LLF	WLF	PBN	Pl yd	Infic	LO	PAS	GRC
12	3.5	9	4	4	0.9	13	12	So	A	3	db
13	11	9.5	3.5	9.0	1.0	5	28	Lo	A	4	lb
14	10.0	11.0	4.0	6.0	1	14.0	18	tc	A	4	rb
15	17.5	8.5	3.5	6.5	1.0	8	40	Mo	A	3	db
16	4	6.5	4	5.5	0.9	12	20	So	B	4	lb
17	16	9.5	4	7.5	1	7	34	Mo	A	3	Rb
18	1.0	6.0	4.5	6	1	9	28	So	A	3	rgb
19	16	10	4	7	1.1	8	12	Mo	A	4	db
20	8.0	8.5	4.5	9.5	1	6.0	48	Mo	A	4	rb
21	12.0	6.5	3.5	4.0	1.2	7.0	26	tc	A	3	rgb
22	12.0	6.5	3.0	6.0	0.9	8	26	I	A	3	lb
23	4.0	5.5	4.0	5.0	0.9	9	18	I	B	4	db
24	20.0	11.5	4.5	5.0	0.9	7.0	32	I	A	4	lb
25	12.0	7.5	6.5	5.0	1	10.0	16	So	A	5	rb
26	12	11	4	9.5	1.0	7	48	Tc	A	4	Rgb
27	10.0	7.5	4.5	4.5	1	8.0	36	O	A	4	db
28	8.0	5.0	3.0	4.0	0.8	6	15	tc	A	4	db
29	17	6	3.5	5.5	1.0	6	63	tc	A	2	db
30	10	17	3	16.5	0.7	5	42	Lo	A	3	Rb

**Table 5** List of accessions used in the current study

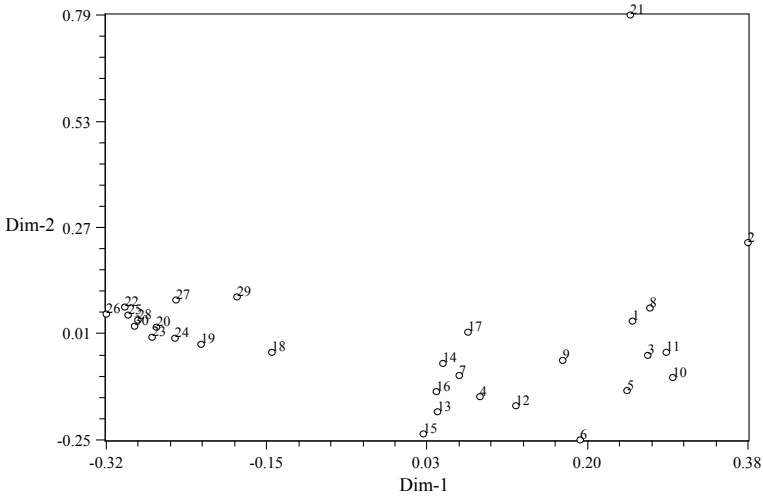
S. No.	IE No.	Origin
1	2457	Kenya
2	2871	Zambia
3	3392	Zimbabwe
4	4757	India
5	3317	Zimbabwe
6	3165	Zambia
7	4734	India
8	6165	Nepal
9	3391	Zimbabwe
10	4497	Zimbabwe
11	5870	Nepal
12	6082	Nepal
13	4565	Zimbabwe
14	6154	Nepal
15	6337	Zimbabwe
16	6350	Zimbabwe
17	7018	Kenya
18	4121	Uganda
19	4491	Zimbabwe
20	3614	Unknown
21	2034	India
22	3470	India
23	6221	Nepal
24	3721	Uganda
25	501	India
26	5201	India
27	5817	Nepal
28	2093	India
29	3475	India
30	2790	Malawi

## 4 Discussion

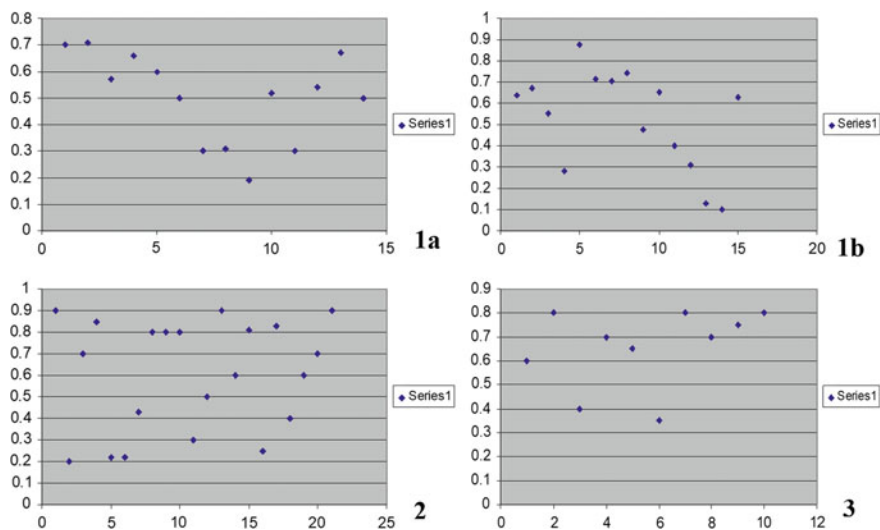
The difference in the clones having SSRs and the ones found suitable for developing markers is explained by several factors. The previous report by [4] revealed the percentage of clones having SSRs varied with RE digestion (1.95, 1.80%, 2.57, for *HindIII*, *PstI* and *SalI* respectively); Clones having multiple SSR motifs were counted once only. Some SSRs might have been present in the un-sequenced clones



**Fig. 1** Dendrogram showing clustering of 30 Finger millet accessions based on UPGMA. Dendrogram was constructed using NTSyS-pc 2.0



**Fig. 2** Plot of genetic distances (PCoA Plot) among 30 Finger millet accessions based on multidimensional scaling analysis. PCoA plot for the second versus first axes were estimated for combined data from markers on thirty finger millet accessions. Lanes 1–30: Finger millet accessions 1–30 as given in Table 5



**Fig. 3** Scatter of PSI values with different markers. **1a** scatter of values for SSR primers developed by us. **1b** scatter of values for SSR primers taken from previously reported data. **2**: scatter of values for EST-SSR primers. **3**: scatter of values for ISSR primers. Figure is drawn taking all the PSI values of the primers. X-axis represents number of primers, Y-axis represents PSI value

as all the clones were not sent for sequencing. The percentage of clones sequenced to the number of primer designed based on the flanking sequences varied. Gao et al. [25] identified only 14 unique SSR sequences in 256 clones. Though 50 plus clones had SSRs in them only 20 were used for primer designing in this study. Only the primers with above 10 repeats were taken as primers; shorter repeats usually produced monomorphic PCR products [26].

EST-SSR markers developed by data mining involve low cost and time [11, 27]. Recent studies report that in several plant species the number of microsatellites are much higher in ESTs than in genomic DNA [12]. EST-SSRs could be of great value in cross species studies and also filling the existing linkage maps. 33 EST-SSRs are used in present study based on their high repeat number [26].

ISSR markers identified based on GA and AG repeats are currently used in the present study as they were reported to be informative in studying genetic relationships among diverse rice germplasm [28]. GA repeats are reported to control gene expression in plants and animals [29, 30]. The bands exhibited by an ISSR marker with a specific microsatellite repeat gives the estimation of motif abundance in the genome [31]. The presence of higher bands only when  $(CT)_n$ ,  $(GA)_n$  based primers are used indicates the fact that CT and GA repeats are frequently present in the Finger millet genome. Usage of ISSR markers to determine diversity among Finger millet accessions has not been previously well investigated as done in the current study.

With increasing number of varietal collections and elite varieties, the usage of molecular markers for genetic selection will be a good option. 83 primers were used

in study out of which 29 SSR markers, 21 EST-SSR markers, 10 ISSR markers were found to be polymorphism. Based on the scores the accessions are grouped as four classes. The markers identified in the current study are found to be highly polymorphic and could be used in further studies like MAS.

Molecular marker technologies when complemented with phenotyping data can facilitate the crop improvement schemes; for population structure studies, identifying agronomic and disease resistance trait QTLs. The reliability and accuracy of various markers in identifying the genetic variation present in between different genotypes/varieties has speeded up the varietal selection process which is usually time taking and laborious. Different SSRs are currently being used for genomic studies and inter and intra specific studies in monocots as well as dicots [32–34]. There is a need for large number of highly variable molecular markers for different research/breeding programs. The markers identified here could be used for pre-screening of existing and elite finger millet varieties and breeding programs for introgression of important traits such as blast disease resistance, drought tolerance, lodging etc.

**Conflict of interest statement** The authors have no competing interests.

## References

1. W.E. Barbeau, K.W. Hilu, Protein, calcium, iron and amino acid content of selected wild and domesticated cultivars of finger millet. *Plant Foods Hum. Nutr.* **43**, 97–104 (1993)
2. A.S. Vadivoo, R. Joseph, N.M. Ganesan, Genetic variability and diversity for protein and calcium contents in finger millet (*Eleusine coracana* (L.) Gaertn) in relation to grain colour. *Plant Foods Hum. Nutr.* **52**, 353–364 (1998)
3. D.R. Fernandez, D.J. Vanderjagt, M. Millson, Y.S. Huang, L.T. Chuang, A. Pastuszyn, R.H. Glew, Fatty acid, amino acid and trace mineral composition of *Eleusine coracana* (Pwana) seeds from northern Nigeria. *Plant Foods Hum. Nutr.* **58**, 1–10 (2003)
4. M.M. Dida, S. Ramakrishnan, J.L. Bennetzen, M.D. Gale, K.M. Devos, The genetic map of finger millet, *Eleusine coracana*. *Theor. Appl. Genet.* **114**(2), 321–332 (2007)
5. L. Arya, M. Verma, V.K. Gupta, J.L. Karihaloo, Development of EST-SSRs in finger millet (*Eleusine coracana* ssp *coracana*) and their transferability to pearl millet (*Pennisetum glaucum*). *J. Plant Biochem. Biotechnol.* **18**, 97–100 (2009)
6. D. Gimode, D.A. Odeny, E.P. de Villiers, S. Wanyonyi, M.M. Dida, E.E. Mneney, Identification of SNP and SSR markers in finger millet using next generation sequencing technologies. *PLoS ONE* **11**(7), e0159437 (2016)
7. I.N.B.L. Reddy, D.S. Reddy, V.P. Reddy, M.L. Narasu, S. Sivaramakrishnan, Efficient microsatellite enrichment in finger millet (*Eleusine coracana* (L.) Gaertn)—an improved procedure to develop microsatellite markers. *Asian Australas. J. Plant Sci. Biotechnol.* **5**(1), 47–51 (2011)
8. I.N.B.L. Reddy, S. Sivaramakrishnan, Identification of SSR markers which could differentiate blast disease resistance accessions in finger millet (*Eleusine coracana* (L.) Gaertn.). *J. Crop Sci. Biotechnol.* **20**, 37 (2017)
9. P. Panwar, M. Nath, V.K. Yadav, A. Kumar, Comparative evaluation of genetic diversity using RAPD, SSR and cytochrome P450 gene based markers with respect to calcium content in finger millet (*Eleusine coracana* L. Gaertn.). *J. Genet.* **89**, 121–133 (2010)

10. H.D. Upadhyaya, C.L.L. Gowda, D.S.S.S.R. Sastry, Plant genetic resources management: collection, characterization, conservation and utilization. *J. SAT Agri. Res.* **6**, 16 (2008)
11. R.V. Kantety, M. La Rota, D.E. Matthews, M.E. Sorrells, Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.* **48**, 501–510 (2002)
12. M. Morgante, M. Hanafey, W. Powell, Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194–200 (2002)
13. G. Toth, Z. Gaspari, J. Jurka, Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967–981 (2000)
14. O.N. Obidiegwu, H. Parzies, J.E. Obidiegwu, Development and genotyping potentials of EST-SSRs in finger millet (*E. Coracana* (L.) Gaertn.). *Int. J. Genet. Genomics.* **2**(3), 42–46 (2014)
15. B.L.I.N. Reddy, N.M. Lakshmi, S. Sivaramakrishnan, Identification and characterization of EST–SSRs in finger millet (*Eleusine coracana* (L.) Gaertn.). *J. Crop Sci. Biotech.* **15**(1), 9–16 (2012)
16. E. Zietkiewicz, A. Rafalski, D. Labuda, Genome fingerprinting by simple sequence repeat (SSR)—anchored polymerase chain reaction amplification. *Genomics* **20**, 176–183 (1994)
17. B. Bornet, M. Branchard, Nonanchored inter simple sequence repeat (ISSR) markers: reproducible and specific tools for genome fingerprinting. *Plant Mol. Biol. Rep.* **19**, 209–215 (2001)
18. R. Prabhu, N.M. Ganesan, Genetic diversity studies in ragi (*Eleusine coracana* (L.) Gaertn.) with SSR and ISSR markers. *Mol. Plant Breed.* **4**, 141–145 (2013)
19. R. Gupta, K. Verma, D.C. Joshi, D. Yadav, M. Singh, Assessment of genetic relatedness among three varieties of finger millet with variable seed coat color using RAPD and ISSR markers. *Genet. Eng. Biotech. J.* **2**, 1–9 (2010)
20. J. Sambrook, E.F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual*, 2nd edn. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989)
21. D. Milbourne, Comparison of PCR based marker systems for the analysis of genetic relationship in cultivated potato. *Mol. Breed.* **3**, 127–136 (1997)
22. F.J., Rohlf, *NTSYS-pc version 2.0* (State University of New York, Exeter Software, Setauket, New York, 1993)
23. J.C. Gower, Some distance properties of latent root and vector methods used in multivariate data analysis. *Biometrika* **53**, 315–328 (1966)
24. P.H.A. Sneath, R.R. Sokal, *Numerical Taxonomy—The Principles and Practice of Numerical Classification* (W. H. Freeman, San Francisco, 1973)
25. G.O. Gao, G.H. He, Y.R. Li, Microsatellite enrichment from AFLP fragments by magnetic beads. *Acta Botanica Sinica.* **45**(11), 1266–1269 (2003)
26. X. Qi, S. Lindup, T.S. Pittaway, S. Allouis, M.D. Gale, K.M. Devos, Development of simple sequence repeat markers from bacterial artificial chromosomes without subcloning. *Biotechniques* **31**, 355–362 (2001)
27. W. Powell, M. Morgante, C. Andre, M. Managey, J. Vogel, S. Tingey, A. Rafalski, The utility of RFLP, RAPD, AFLP and SLP (microsatellite) markers for germplasm analysis. *Mol. Breed.* **2**, 225–238 (1996)
28. C.S. Reddy, A.P. Babu, B.P.M. Swamy, K. Kaladhar, N. Sarla, ISSR markers based on GA and AG repeats reveal genetic relationship among rice varieties tolerant to drought, flood, or salinity. *J Zhejiang Univ. Sci. B.* **10**(2), 133–141 (2009)
29. L. Santi, Y. Wang, M.R. Stile, K. Berendzen, D. Wanke, C. Roig, C. Pozzi, K. Muller, J. Muller, W. Rohde, F. Salamini, The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. *Plant J.* **34**(6), 813–826 (2003)
30. B. Van Steensel, J. Delrow, H.J. Bussemaker, Genome wide analysis of Drosophila GAGA factor target genes reveals context dependent DNA binding. *Proc. Nat. Acad. Sci.* **100**(5), 2580–2585 (2003)
31. M.W. Blair, O. Panaud, S.R. McCouch, Inter-simple sequence repeat (ISSR) amplification for analysis of microsatellite motif frequency and fingerprinting in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* **98**, 780–792 (1999)

32. I. Eujayl, M. Sorrells, M. Baum, P. Wolters, W. Powell, Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theor. Appl. Genet.* **104**, 399–407 (2002)
33. P. Rallo, I. Tenzer, C. Gessler, L. Baldoni, G. Dorado, A. Martin, Transferability of olive microsatellite loci across the genus *Olea*. *Theor. Appl. Genet.* **107**, 940–946 (2003)
34. A.L. Westman, S. Kresovich, The potential for cross-taxa simple-sequence repeat (SSR) amplification between *Arabidopsis thaliana* L. and crop brassicas. *Theor. Appl. Genet.* **96**, 272–281 (1998)

# Social Media—Impact on Sexual and Reproductive Knowledge of Adolescents in South India



N. Rajani and A. Akhila

**Abstract** The study reveals social media impact on sexual reproductive knowledge of adolescents' in Tirupati, Chittoor district of Andhra Pradesh. The investigator adopted descriptive survey design to collect the data from 525 boys and girls in the age group of 10–19 years. Multistage random sampling method was used to select the respondents for the study. Appropriate statistical methods were used to analyse the data. The data revealed that there is significant influence of socio-demographic characteristics on respondents' usage of SM. Boys were using the SM actively than girls in the age group of boys 10–13 and 13–16 years of age were using social media actively than girls. Adolescents across sexes was found to use the SM especially during the 16–19 years. Data from the study clearly indicates that adolescents being active SM users need to be properly sensitized towards sexual health and reproduction. So that they emerge as healthy human beings.

**Keywords** Adolescent · Social media · Reproductive health · Knowledge

## 1 Introduction

Health is of prime concern across the entire life cycle of the human being. Adolescent is being the stage where the child blooms to be a full fledged man/woman is a very critical age in the human life cycle. After all the child is the father of man. The current scenario with explosive SM concerns have a very serious impact on the adolescents' knowledge, Attitude and Behaviour. During adolescence, besides pubertal changes, mental health, emotional health and physiological changes of the body create a gamut of emotions. The adolescent is sometime happy, angry, aggressive, withdrawn and at times lacks confidence. At this juncture he/she cannot totally comprehend bodily changes and cope with hormonal changes within. The body may express a particular

---

N. Rajani · A. Akhila (✉)  
Department of Home Science, SPMVV, Tirupati, India  
e-mail: [akhidarlg786@gmail.com](mailto:akhidarlg786@gmail.com)

N. Rajani  
e-mail: [rajani.nallanagula@gmail.com](mailto:rajani.nallanagula@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_28](https://doi.org/10.1007/978-3-030-46939-9_28)



desire but the mind and social expectations may give conflicting instructions. Due to the lack of sufficient social exposure and perhaps inadequate familial support the adolescent turns to the SM seeking answers/comfort on the internet.

Many adolescents are addicted to SM and globally about 3.77 billion people are using the internet and other SM websites. SM is becoming integral part of the day-to-day lives of majority in India [1–3]. Cannot imagine their life without Facebook, You tube, Instagram, whatsapp. The web and other new media advances have made sexually explicitly materials more available to youngsters than any other time [4]. New media technologies are making sexuality explicit materials accessible to youngsters than yesteryears [5]. Media making adolescents more active in sexual behavior than earlier. The family gives a set of socially acceptable, behavioral norms where as the digital media and peer exposure may give feedback and responses which may be quite opposite towards the family had expected. Inability to cope with these conflicts the adolescents sexual health and concept of sexual reproduction may get biased and sometime incorrect exposure. Very many studies have reported on the various aspects of adolescents interaction with sexual health, and his own body as also reproductive health [6]. Against this background outlined above the current study was taken to gain insights into south Indian adolescents' knowledge concerns and understanding of sexual health, reproduction and sexual behavior.

More than  $\frac{1}{4}$  of the world's population is between the ages of 10-24 and eighty four living in less developed countries. The procreative and sexual health choices they create nowadays can have an effect on the health and well being of their communities and of their countries for many years to come back. Applicable steps created by the adolescents is AN intergenerational investment with immense advantages to ulterior generations.

The main objective of the study was to see the impact of socio-demographic profiles on Sexual and Reproductive Health Knowledge (SRHK) and usage of SM.

## 2 Methodology

In the present study the investigator felt that the descriptive survey was the evidence to substantiate the objectives of the study. The topic for the investigation selected was Sexual and Reproductive Knowledge among adolescents. The study adopted multi-stage sampling procedure. The locale of the study was Tirupati, Chittoor district of Andhra Pradesh State. Tirupati is a educational hub and a large number of adolescents belonging to entire state were studying in the town. Hence Tirupati town was selected purposively.

A tool which would provide sound data is required to collect accurate information from the subjects included in the study. Therefore a well delineated questionnaire was developed which includes questions related to personal details, details of SM usage, and knowledge of adolescents' on sexuality and reproductive health. Knowledge of adolescents' was collected using a rating scale. The tool was given to the experts of

concern field to assess the suitability language, content and domain. Based on the experts suggestions the tool was modified.

The sample consists of 525 boys and girls in the age group of 10–19 years. The inclusion criteria for the study was the subjects who attained puberty and willingness. Consent was taken from the school and college authorities and also from the participants. While conducting the study the investigator felt that the issue being of a sensitive nature, the subjects were told that total confidentiality would maintained. This was because of the study ethics involved. Appropriate statistical analysis were used, so as to get substantial results.

### 3 Results and Discussions

For any study the basic background of socio-demographic features of the respondents is required, therefore in the current study too, age, sex, education level, economic status and type of family were included to how the impact of these variables on the respondents SRHK and usage of SM.

While examining the distribution of sample across three age group as given in Table 1 there was almost equal distribution of respondents in 13–16 years (40.5%) and 16–19 years (37.3%) age group but only 21.9% were 10–13 years of age group. Similarly

**Table 1** Respondents distribution as per socio-demographic profiles

Category	Number	Percentage
<i>(1) Age in years</i>		
• 10–13	115	21.9
• 13–16	213	40.5
• 16–19	196	37.3
<i>(2) Sex</i>		
• Male	261	49.7
• Female	264	50.2
<i>(3) Educational level</i>		
• Middle school	98	18.6
• High school	208	39.6
• College going	219	41.7
<i>(4) Type of family</i>		
• Nuclear	347	66.09
• Joint	178	33.9
<i>(5) Economic status</i>		
• Low	92	17.5
• Middle	169	32.1
• High	264	50.2

on examining the sample distribution both sexes was almost equitable. Among the respondents 41.7% were college goers, 39.6% high school, 89.6% were in middle school.

Family composition from the study has clearly established a majority of the families belonging to the nuclear family type (66.09%). This also could probably be attributed to the urban influence. The economic status of the respondents has established that about 50.2% belonged to high income group and middle income group 32.1%, where as 17.5% of the respondents were from the lower economic status. From this study data it was observed that a larger percent belong to the high and middle class. The probable reason for this could be that the sample belonged to the urban areas. This being so the parents had better employment opportunities and earned relatively better incomes.

Influence of SM among respondents was examined in depth. There was a predominant use of SM in 16–19 (37.3%) the age group and 13–16 (32.71%) age group when compared to 10–13 years (15.04%) age group. Further the 16–19 age group was observed to have higher usage of the SM (37.3%) and observed in particular boys of this age group. The reasons for this kind of data representation was—girls having limited access to SM. Also boys had easier access because they could freely move out of the house and visit internet cafes when compared to the girls.

The type of social media used by the respondents were—You tube (100%), Whatsapp (67.80%), Facebook (41.7%), Twitter (22.47%), Others like Instagram, Myspace (18.6%). Youtube, Whatsapp, facebook, are used more likely than the other SM. Instagram, Twitter and Myspace is less likely used among all respondents. All these are the applications of internet based that helps the youngsters to receive their friendship, to make new friends, exchange of ideas, images, videos and to know the information which is not discussed openly (6). SM has become deeply ingrained in daily lives of youngsters, they are addicted to checking their favorite SM platform. In the present study a high percent (52.5%) of respondents on an average spending 2–4 h/day, 21.7% of them were spending 4–6 h/day, 25.7% clocking in at about 2 h/day.

The data in Table 2 indicates more than half of the respondents 51.2% using social media for movies/music/pictures; 29.9% were searching media for sexual and reproductive information.

The data in Table 3 reveals that in the age group of 10–13 the knowledge levels were comparatively low where as in 13–16 years and 16–19 years age groups the knowledge levels were more towards the high levels on being subjected to chi-square test. Knowledge levels of the 13–16, 16–19 years age groups was significantly high when compared to 10–13 years age group. The male (80.8%) respondents in the study had a high level SRHK knowledge when compared to the females who had moderate knowledge (73.1%). Even this value was significant at ( $\chi^2$ : 307.3, Df: 2, *P* value: 0.0000). The Education levels were along expected lines with high school (74.3%) and college going students (84.9%) having better SRH knowledge. The economic status and type of family had significant impact on the SRH knowledge of the respondents. Respondents from nuclear families (85.1%) were found to have expressed greater knowledge when compared to the joint families (21.9%). A similar

**Table 2** Percentage distribution of respondents' social media usage

Social media usage (SM)	Number	Percentage
<i>(1) Age at first use of SM (In years)</i>		
• 10–13	79	15.04
• 13–16	172	32.71
• 16–19	196	37.3
<i>(2) SM use most</i>		
Face book	219	41.
Twitter	118	22.47
Watsapp	356	67.80
YouTube	525	100
Others (Instagram, MySpace)	98	18.6
<i>(3) Average time spent on SM (h)</i>		
• 0–2	135	25.7
• 2–4	276	52.5
• 4–6	114	21.7
<i>(4) SM used for</i>		
For School work	6	1.14
To Meet friends/People	226	43.04
To Update/Upload pictures	138	26.2
For Knowledge purpose	58	11.04
To Watch videos/Music	97	18.4
<i>(5) Activity enjoyed by SM</i>		
Non sex chat	23	4.3
Sexual &Reproductive	157	29.9
Movies/music/pictures	269	51.2
School/college work	57	10.8
Others	19	6.2

pattern was observed in the economic status of the respondents too. The respondents from high and middle class had better knowledge (Table 4).

The years of usage of SM when examined against the socio-demographic variables also reiterated the findings of the SRH knowledge levels among the respondents. The respondents particularly boys ( $\chi^2$ : 76.066, Df: 2,  $P$  value: 0.000) and respondents belonging to high school and colleges as also high economic status were found to have greater access and using more years of the varied SM, hence their SRHK levels were better.

On examination of the usage of SM for the number of years by the respondents irrespective of age, sex, educational levels, type of family and also economic status, it was observed that the usage of SM was around 4 years only. However, the male respondents (35%) and college students as also the high economic status families did indicate their usage of SM was more than 5 years.

The data in Table 5 also indicates the more number hours used by the respondents

**Table 3** SRH Knowledge of respondents/socio demographic status

Category	Low		Moderate		High		$\chi^2$	Df	P value
	No	%	No	%	No	%			
<i>(1) Age</i>									
10-13	75	65.21	31	26.9	9	7.82	407.139	4	0.0000
13-16	11	5.16	14	67.13	59	27.6			
16-19	5	2.5	22	11.22	169	86.2			
<i>(2) Sex</i>									
Male	21	8.04	29	11.1	211	80.8	307.03	2	0.0000
Female	56	21.2	193	73.1	15	5.68			
<i>(3) Educational level</i>									
Middle	65	66.3	26	26.5	7	7.1			
High school	7	3.3	47	22.5	154	74.3	297.151	4	0.0000
College	4	1.82	29	13.2	186	84.9			
<i>(4) Type of family</i>									
Nuclear	14	4.03	38	10.9	295	85.1	297.054	2	0.0000
Joint	39	21.9	127	71.3	12	6.7			
<i>(5) Economic status</i>									
Low	75	81.5	13	14.1	4	4.3			
Middle	6	3.5	28	16.5	135	79.8	304.179	4	0.0000
High	17	6.4	69	26.1	178	67.4			

**Table 4** Years of SM usage of respondents/socio-demographic status

Years of usage		<2 years		4 years		>5 years		$\chi^2$	Df	P value
		No	%	No	%	No	%			
<i>(1) Age</i>										
10-13		104	90.4	11	9.5	0	0	180.021	4	0.0000
13-16		126	59.15	79	37.08	8	3.7			
16-19		29	14.79	149	76.02	18	8.6			
<i>(2) Sex</i>										
Male		34	13.02	135	51.7	92	35.2	76.066	2	0.0000
Female		81	30.6	167	63.2	16	6.06			
<i>(3) Educational level</i>										
Middle		93	93.8	5	5.1	0	0			
High school		35	16.8	157	75.4	16	7.6			
College		9	4.1	132	60.2	78	65.6	353.369	4	0.0000
<i>(4) Type of family</i>										
Nuclear		32	9.2	236	68.01	79	22.7	131.979	2	0.0000
Joint		94	52.8	78	43.8	6	3.3			
<i>(5) Economic status</i>										
Low		65	70.6	25	27.1	2	2.17			
Middle		20	11.8	114	67.4	35	20.7	240.335	4	0.0000
High		13	4.9	123	46.5	128	48.4			

**Table 5** Hours of SM usage of respondents/socio-demographic status

Hours of usage		<2 years		4 years		>5 years		$\chi^2$	Df	P value
		No	%	No	%	No	%			
<i>(1) Age</i>										
10-13	98	85.2	15	13.04	2	1.7	449.731	4	0.0000	
13-16	14	6.5	117	54.9	82	38.4				
16-19	3	1.5	28	14.2	165	84.1				
<i>(2) Sex</i>										
Male	2	0.7	21	8.04	238	91.1	258.555	2	0.0000	
Female	9	3.4	198	75	57	21.5				
<i>(3) Educational level</i>										
Middle	95	96.9	3	3.06	0	0				
High school	27	12.9	132	63.4	49	23.5	542.413	4	0.0000	
College	4	1.8	26	11.8	189	86.3				
<i>(4) Type of family</i>										
Nuclear	62	17.8	197	56.7	88	25.3	29.840	2	0.0000	
Joint	58	32.5	106	59.5	14	7.8				
<i>(5) Economic status</i>										
Low	43	46.5	25	27.1	24	26.0				
Middle	9	5.3	134	79.2	26	15.3	267.020	4	0.0000	
High	5	1.8	81	30.6	178	67.4				

16–19 years age group (84.1%). The probable reason for this age group of respondents spending more hours on the SM could be due to greater accessibility to the media when compared to the other age groups. The respondents 10–13 years age group (85.2%) were used SM on average 2 h/day. May be restrictions on usage of internet, more academic work and parents supervision in this age group made them utilized SM less number of hours.

## 4 Conclusion

In order to study the effect of SM on the knowledge of SRH by the adolescents group the socio-demographic variables, access to SM and the time spent on the SM was studied. In the socio-demographic variables adolescents belonging to high school and college group and having nuclear family were found to spend more time using the SM. Hence their knowledge of SRH was significantly better than younger age group (10–13 years). The socio-economic status also having a direct impact on access to media and time spent, therefore adolescents' belonged to the high and medium socio-economic class were better informed about SRHK related issues. Since the boys had the advantage of going out to visit internet cafes, using their friends mobiles, they were found to have more about SRHK. The study can safely conclude that economic status, higher levels of education levels and nuclear families does have an association with greater knowledge of SRHK.

## References

1. ICT Facts and Figures. Last accessed on 31 Aug 2017. Available from: <http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx>. (2017)
2. Digital in: Global Overview—We Are Social. Last accessed on 18 Sep 2017. Available from: <https://www.wearesocial.com/special-reports/digital-in-2017-global-overview> (2017)
3. India Has 462 Million Internet Users; 79% Traffic is Mobile. Last accessed on 24 Sep 2017. Available from: <https://www.techinasia.com/india-462-million-internet-users-79-traffic-mobile> (2017)
4. N.R. Masthi, S.R. Cadabam, S. Sonakshi, Facebook addiction among health university students in Bengaluru. *Int. J. Health Allied Sci.* **4**(1), 18 (2015)
5. S.H. Subba, C. Mandelia, V. Pathak, D. Reddy, A. Goel, A. Tayal et al., Ringxiety and the mobile phone usage pattern among the students of a medical college in South India. *J. Clin. Diagn. Res.* **7**, 205–209 (2013)
6. A. Lenhart, K. Purcell, A. Smith, K. Zinuhr, *Social Media Mobile Internet Use Among Teens and Young Youth Adults Washington DC Pew Internet and American Life Project*. [www.pewinternet.org](http://www.pewinternet.org). Assessed on 28 Dec 2014



# Wearable Electronic Gloves in Two-Way Communication to Convert Signs into Speech



S. Swarnalatha, Anusha Manubrolu, and Pooja Dande

**Abstract** Usual tasks such as, communicating messages and expressing feelings/thoughts, are easy for an average person with all functional organs. However, for differently abled persons day-to-day conversations are challenging. So they prefer writing down messages or making signs to communicate. The major drawback of signs is that, they can only be understood by those who are aware of the sign language. In this project, a two way communication system is introduced to convert signs to audio output using electronic gloves and speech to text using a speech-to-text converting application. Flex sensors are used to capture signs by the moments of fingers and their output is given as input to microcontroller and microcontroller's output to speaker. Output from software which converts speech into text is given as input to microcontroller with the help of Bluetooth and LCD is used to display text from speech-to-text converter.

**Keywords** Signs · Flex sensors · Speech-to-text converter · Bluetooth

## 1 Introduction

In world's population 5% of people has hearing disability and for the others who engage with them communication is not easy. Listening and speech disability in people causes a decrease in ratio of literate and employment. Signs can be used to communicate which uses gestures and postures to convey a message.

There are different sign languages in different countries and regions. Like spoken languages, sign languages also differ from each other based on regions. Hand gestures

---

S. Swarnalatha · A. Manubrolu (✉) · P. Dande  
Department of Electronics and Communication Engineering, S.V. University College of Engineering, Tirupati, AP, India  
e-mail: [anushamanubrolu@gmail.com](mailto:anushamanubrolu@gmail.com)

S. Swarnalatha  
e-mail: [swarnasvu09@gmail.com](mailto:swarnasvu09@gmail.com)

P. Dande  
e-mail: [dandepooja438@gmail.com](mailto:dandepooja438@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_29](https://doi.org/10.1007/978-3-030-46939-9_29)



**Fig. 1** Sign language symbols. Link: <https://english-speak-english.com/body-language-gestures-hand-signals-sign-language/?lang=de>

along with face expressions forms signs, where as in this project only hand gestures are captured by deploying flex sensors.

In Electronic gloves flex sensors plays a major role. Each finger has a flex sensor, resistance of the flex sensors varies in accordance with the deflection/bending of the finger. Output from flex sensors are processed in atmega328 and output fed to a text-to-speech converter (Fig. 1).

## 2 Earlier Work

The conventional idea for gesture recognition is to use a camera for capturing gestures. It is less user friendly because if we use it in area where more people gathered, too many gestures may be captured from the people around. Then it becomes difficult to remove those unwanted gestures from the wanted one and also it is difficult to carry around.

One way communication providing sign language interpreter is also available. Using this interpreter signs are understood by normal person but differently abled cannot understand the message from normal persons. Normal person cannot reply back due to lack of knowledge of sign language.

The purpose of proposed project is to provide a means to basic communication and makes two-way communication possible. It takes hand gestures from people with hearing or listening disability as input and maps the gesture with corresponding

text and gives audio as output with the help of a converter. To make two-way communication possible a software is needed which takes speech and gives corresponding text output. It takes speech from normal person as input and gives text as output and then its output fed to LCD.

### 3 Hardware Description

#### (a) Flex sensors

Flex sensors resistance varies with its deflection so it is also named as variable resistor. These are normally attached to surface like fingers whose deflection/bent is to be quantified. These are used as door sensors to know whether the door is closed or not based on the bend resistance of Flex sensor. Flex sensor is a device with two terminals but like diode polarized terminals are not present. So there is no negative and positive. They can operate on low voltages, flat resistance is 25 KΩ. If flex sensor is bent 45°, then resistance becomes double than before and when bent 90°, resistance increases to four times than the nominal resistance (Fig. 2).

For our convenience we convert resistance into voltage using a voltage divider circuit (Fig. 3).

$$V_o = V_{cc} (RV1 / (R1 + RV1))$$

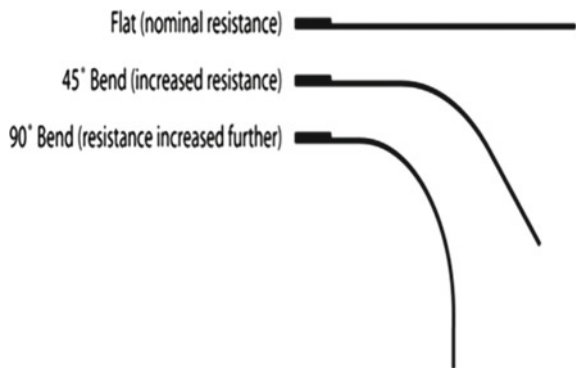
RV1-Flex sensor resistance

Voltage  $V_o$  increases when bent in flex sensor is more and vice-versa.

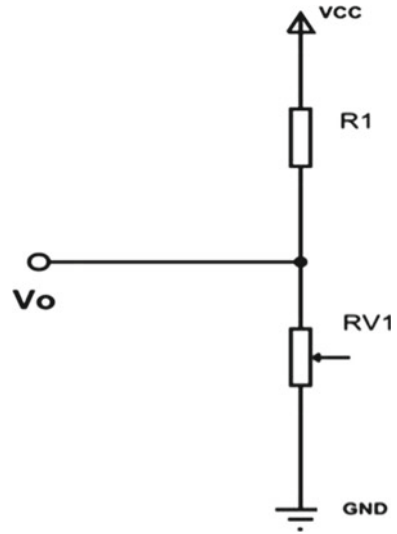
#### (b) ATmega328

ATmega328 belongs to AVR family microcontroller and has 28 pins. It supports data up to 8 bits. It has 32 KB internal built-in memory. It has 3 built-in timers two are 8-bit and one is 16-bit timer. It dissipates low power, cost effective and has programming lock for security purposes.

**Fig. 2** Flex sensor variable resistance reading. Link: <https://components101.com/sensors/flex-sensor-working-circuit-datasheet>



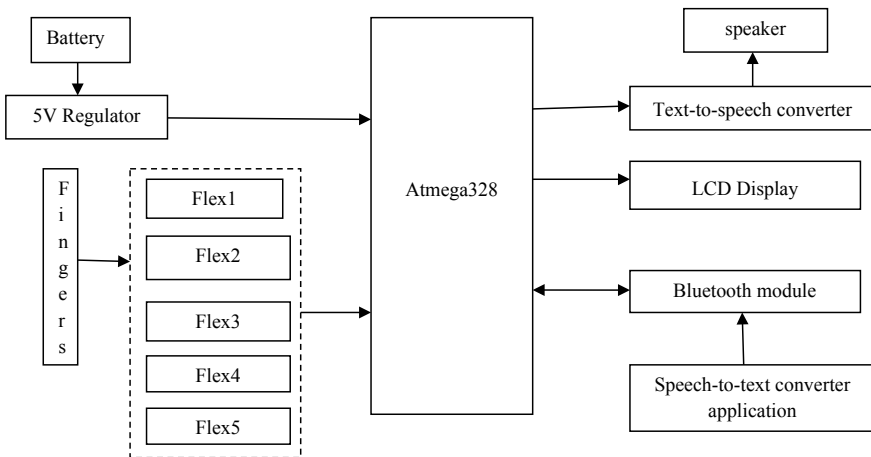
**Fig. 3** Voltage divider circuit. Link: <https://components101.com/sensors/flex-sensor-working-circuit-datasheet>



ATmega328 has an adc converter which is built-in converts analog voltage into digital form (Fig. 4).

**(c) Text to speech conversion module**

It contains 4 interface/power pins. TX-OUT is typically connected to RXD pin of given microcontroller. We can connect RX-IN pin to microcontrollers TXD pin if data source is microcontroller. If the gloves output is matched with pre-defined data then microcontroller gives input to text to speech converter.



**Fig. 4** Block diagram

#### (d) **LCD Display**

LCD is used for displaying text coming from the Bluetooth module. We can display numbers, letters and graphics. LCD has 8-bit data line where it can be made to work in 4-bit mode or 8-bit mode. Three control pins are there RS, EN and RW. To write RW pin should be low and to read it should be high. Using commands we can make letters to change their displaying position.

#### (e) **Bluetooth module**

HC-05 is a Bluetooth serial port protocol module and easy to use, designed for transparent serial wireless connection. We can use it in Master/Slave configuration. The role of module configured only by AT commands. Default baud rate of HC-05 is 9600 bps in data mode and 38,400 bps in command mode.

## **4 Methodology**

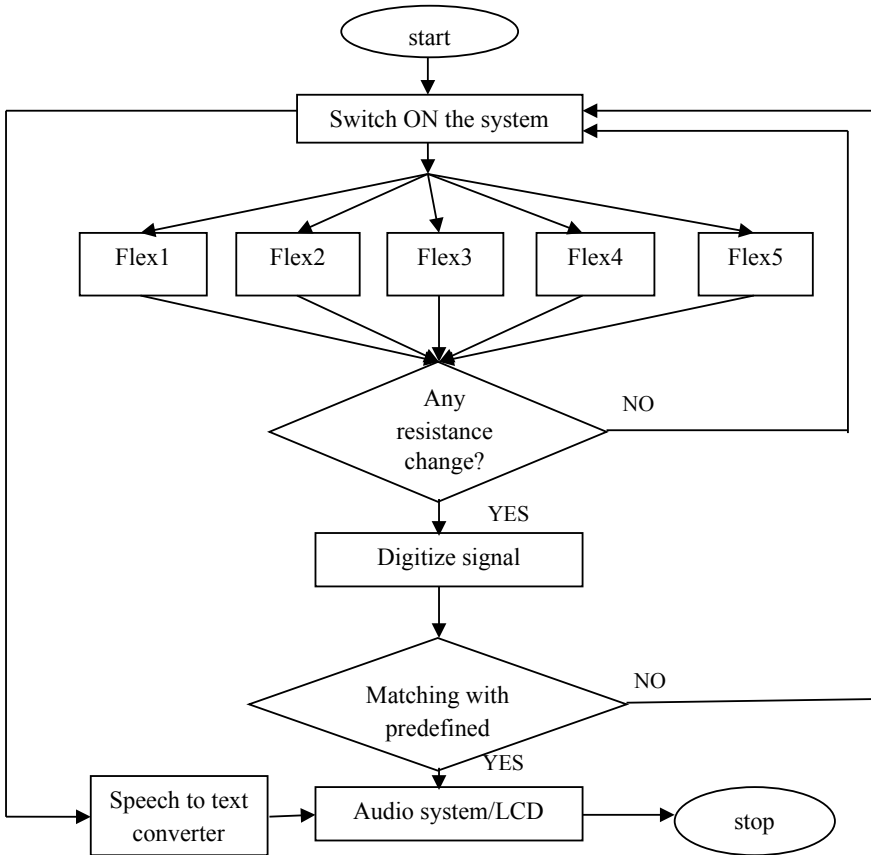
This study aims at the development of a two-way communication system to convert signs by analyzing gestures using a smart wearable Electronic glove (Fig. 5).

The wearable Electronic glove is designed in order to hold the hardware components and it is flexible with good elasticity. Flex sensors output is given to ATmega328 which is core to the device. When the power is turned ON microcontroller checks for any change in resistance from flex sensors. Upon sensing a change an electrical output will be generated which should be digitized for further processing. ATmega328 has built-in ADC which converts analog signals into digital form and then the output is mapped with predefined data where it matches the signal and transmits the text to text-to-speech converter. Speaker gives the audio output. This is one-way communication.

For two-way communication we need a Speech-to-text conversion application. Speech-to-text conversion application is a kind of software which takes audio and converts it into text. We can also call this application as voice recognition application also. Using a Bluetooth module phone gets paired with the hardware. Through the Bluetooth interface, the output from voice recognition application is given to microcontroller to display in the LCD.

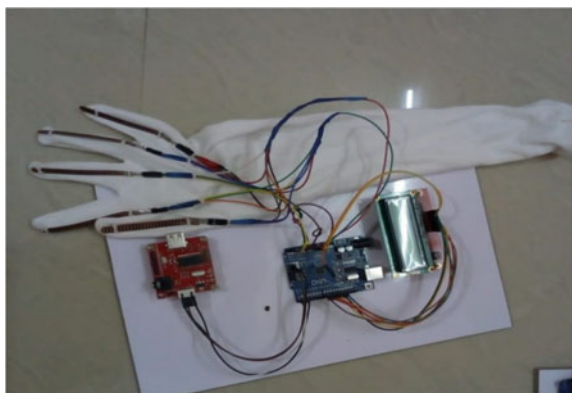
## **5 Result**

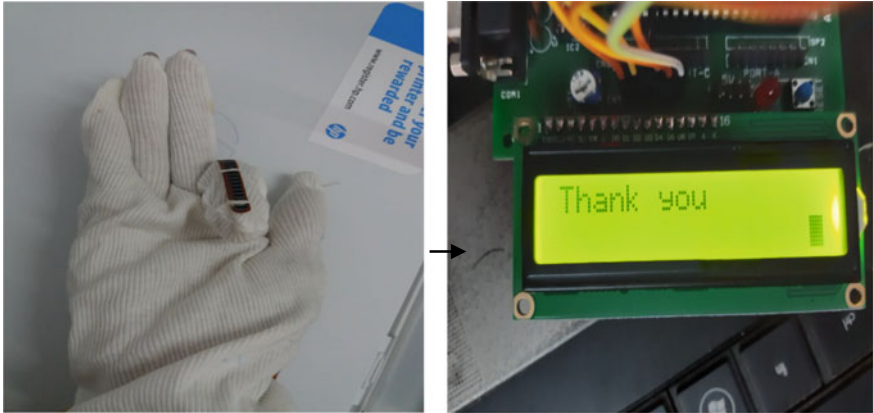
- These gloves converts signs into speech with which deaf and dumb people can express their ideas easily.
- When fingers are bent change in resistance value occurs and predefined data according to the resistance value is sent to the speaker to give audio output. For different resistor values different data fed to speaker (Figs. 6 and 7).



**Fig. 5** Flow chart of two way communication for sign language conversion

**Fig. 6** Wearable electronic glove to convert signs into speech





**Fig. 7** Sign and its corresponding output displayed on LCD screen

- Predefined data is fixed for fixed range of resistances, if resistance value exceeds the range then the data will change.
- For different sentence we need to follow particular signs to get accurate and unique output.
- Using speech-to-text converting app voice get converted into text and sent to microcontroller using Bluetooth to display on LCD screen.

## 6 Conclusion

In this study, a two-way communication to convert signs using wearable electronic gloves system is implemented which is useful for gesture recognition. People with hearing and speech disability can communicate easily with the ones who don't have knowledge of sign language also. Compared to earlier methods present one is easy and cost effective. This system mainly has two parts converting signs to speech and speech-to-text. Bluetooth plays major role to send text from mobile application to microcontroller without wired connections. However, despite of all the progress discussed in this paper, still needs some requirement we can use nano technology to make the project compact.

## References

1. L. Dipietro, A.M. Sabatini, P. Dario, A survey of glove-based systems and their applications. *IEEE Trans Syst. Man Cybern. Part C (Appl. Rev.)* **38**(4), 461–482 (2008)

2. K.C. Sriharipriya, K. Aarthy, T.S. Keerthana, S. Menaga, S. Monisha, Flex sensor based non-specific user hand gesture recognition. *Int. J. Innov. Res. Stud.* **2**(5), 214–220 (2013)
3. V. Shoaib Ahmed, C. Abdul Hakeem, *India submitted the thesis report of Hand Gesture Recognition and Voice Conversion System for Differentially Able Dumb People* (2012)
4. P. S. Havalagi, S. U. Nivedita, The amazing digital gloves that give voice to the voiceless. *Int. J. Adv. Eng. Technol.* **6**(1), 471–480 (2013). ISSN: 2231-1963



# A Perspective Overview on Machine Learning Algorithms



S. Nalini Durga and K. Usha Rani

**Abstract** The latest innovations of technology now a days, avails the massive collection of information from various sources which leads to solve many real life challenges. The key challenge is to design the right model by handling the massive data. From past researches, there were so many misinterpretations made due to wrong choice of models. The lack of skill lies in handling the conservative data, designing the effective reasoning capabilities and handling missing data has triggered an increase in the number of studies using non-conventional methods like machine learning techniques. The study of machine learning which is a subset of Artificial Intelligence focusses on efficient and accurate prediction process by automating the knowledge engineering process avoiding the human intervention. In this research, the focus is made on the popularly used machine learning algorithms like K-Nearest Neighbors (KNN), Naive Bayes (NB), Support Vector Machine (SVM) and Decision Trees (DT) along with their suitability, advantages and disadvantages with performance accuracy.

**Keywords** Machine Learning · Artificial Intelligence · KNN · SVM · Naive Bayes · Decision Trees · Performance Accuracy

## 1 Introduction

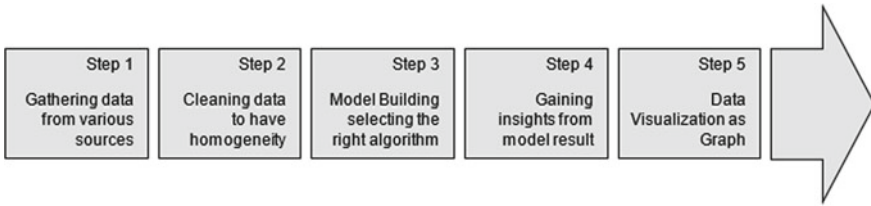
Now-a-days, the leading innovations are designed to reduce intervention of humans with the help of machines and robots by providing the learning mechanism. Artificial Intelligence (AI) plays a key role in such innovations which focus on “training the machine by itself”. The research on AI has come into existence from past years in order to solve many real world challenges. The Machine Learning (ML) is the subset

---

S. Nalini Durga (✉) · K. Usha Rani  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India  
e-mail: [nalini.sst@gmail.com](mailto:nalini.sst@gmail.com)

K. Usha Rani  
e-mail: [usharanikuruba@yahoo.co.in](mailto:usharanikuruba@yahoo.co.in)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_30](https://doi.org/10.1007/978-3-030-46939-9_30)



**Fig. 1** The process of machine learning [1]

of Artificial intelligence, where the computers, software's and devices integrated together to operate successfully by providing the learning or self-training mechanism. The Machine learning revolves around the problem of prediction. With the rapid growth in technologies like Internet of Things (IoT) in the past decade, large volumes of data are collected from various sources in industries like telecom, healthcare, safety and security, finance, retail, transport industries and various forms of patterns were created. The various set of data analytics tools and platforms can be applied on the data collected to identify and interpret the behavioural patterns for solving the real world challenges. The machine learning strategies play a key role in analysing the complex data and identify the hidden patterns, understand market trends and customer preferences. The objective of machine learning research is to make machines to learn and reorganize the patterns on the data collected and with the help of algorithms the new patterns are identified and given as a training set for machine to learn. The process of learning can be referred as a "training" and the output generated from the machine by such training is called "model". The model offers new set of patterns to justify the prediction made. The process involved in machine learning is given in Fig. 1.

Machine learning models define a set of rules with varying amounts of computing power. The more data a machine learning model is served, the more intricate the rules, the more precise the predictions is possible. Many research studies have been conducted to justify the learning process of machines [2, 3] by proposing several approaches and algorithms. The key challenge in machine learning lies in choosing the optimal algorithm for solving the problem with the key parameters like speed, forecast accuracy, training time. In the preceding sections, the major algorithms namely KNN, Naïve Bayes, SVM and Decision Trees are propounded with their set of advantages and disadvantages.

## 2 Types of Learning

The Assessment of machine learning lies in understanding the learning method to perform better on a set of test case taken as an input. The set of Machine learning algorithms and their areas of applications are propounded in the next sections. Figure 2 demonstrates the broad categories of ML algorithms followed by an

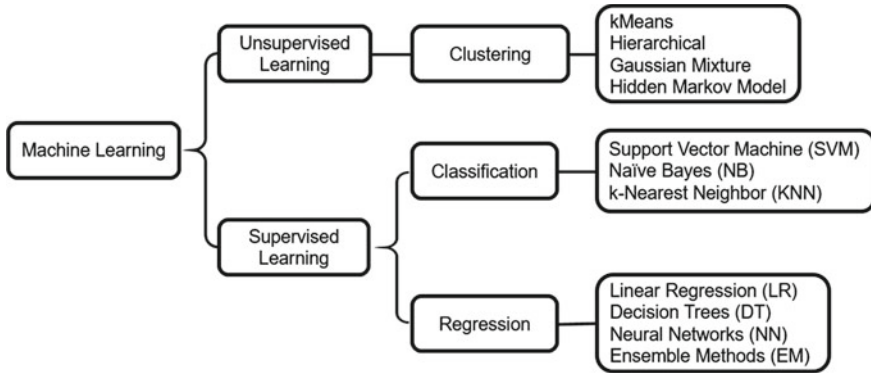


Fig. 2 Classifications of ML algorithms

essential overview along with the respective examples for each category.

The concepts of machine learning is classified into supervised and unsupervised learning techniques which in turn has a sub-classification of Clustering, Classification and Regression techniques with a set of algorithms in each set of sub-classification.

### 2.1 Supervised Learning

The supervised machine learning algorithms are the algorithms which needs external backing. The input dataset is divided and labelled as training dataset and test dataset. The train dataset provides the output variable for forecasting. The algorithms learn the set of patterns from the training dataset after performing the training and verifies with respect to the test data for making approximate or accurate prediction [4]. The workflow of supervised machine learning algorithms is given in Fig. 3. The supervised learning can be implemented in a variety of domains such as marketing, finance and manufacturing.

This model is further classified into *classification* models (classifiers) and *Regression* models where Regression model maps input space into a real-value domain and classifiers forecast the input space into predetermined classes.

### 2.2 Unsupervised Learning

Supervised Machine learning Algorithms act on structured data. On the other hand the unsupervised learning is implemented on unstructured data, unfiltered data. The unsupervised learning does not allow inclination as the range of outcomes is unknown. In this learning, the software is allowed to extract the patterns itself. Clustering and Dimensionality reduction are two major categories falling under this unsupervised

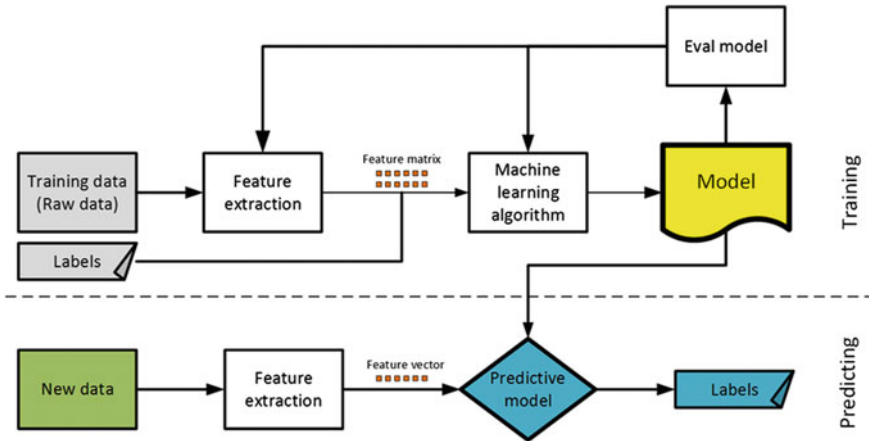


Fig. 3 Workflow of supervised machine learning [5]

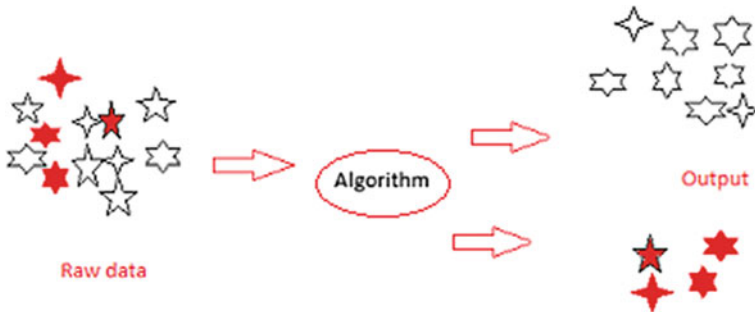


Fig. 4 Outcome of unsupervised learning [7]

learning. Clustering locates to the region based on the location of data points, measures the distance from the centroid of the cluster [6]. This method is used when the  $n$  number of predictions needs to draw from the dataset without any limitations. A sample representation of unsupervised learning is shown in Fig. 4.

### 3 Machine Learning Strategies

The set of Machine Learning strategies that can be followed are dependent on the application considered. For instance, if the application should decide whether the possibility of existence or not, then the decision trees algorithm can be adopted. In this section, some of the popularly used algorithms namely KNN, Naive Bayes, SVM and Decision trees were propounded and it can be generalized that selection

of the algorithm is purely dependent on the application considered for the reset of ML algorithms also.

- (a) *K-Nearest Neighbours*: KNN is a learning by equivalence, where the given test records are compared with the training records for the similarity. The training records are denoted as  $n$  attributes represented in  $n$ -dimensional pattern space. When a new record or test case is given, classifier search for the similar the pattern space in training set records in the form of nearest neighbour mechanism. Figure 5 shows an example.
- (b) *Naïve Bayes*: It is the one algorithm, which depends on conditional probability for classification and clustering tasks. It generates a network namely Bayesian, by creating trees based on their likelihood of happening. It can be implemented for various applications such as Recommendation System and forecasting of cancer relapse or progression after Radiotherapy and text classification industry.
- (c) *Support Vector Machine*: It is a widely used state-of-the-art machine learning technique specifically for classification. The basic principle adopted for SVM is margin calculation. The margins should be drawn, in such a way that the distance between drawn margins should be maximum. Figure 6 shows sample work flow.

Fig. 5 Prediction of voter based on K-Nearest Neighbours [8]

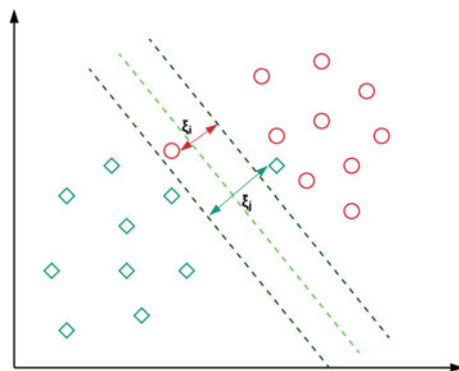
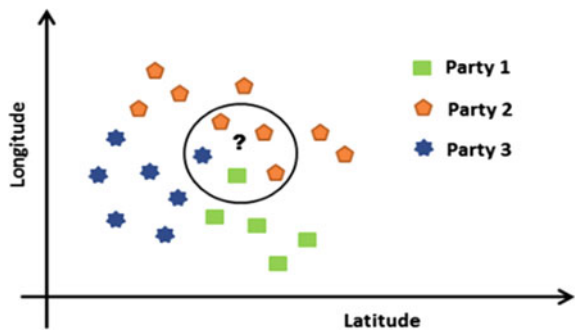
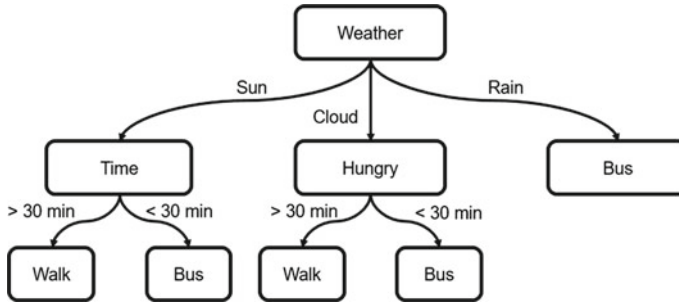


Fig. 6 SVM's soft margin formulation technique in action [9]



**Fig. 7** Predicting how to make the journey using decision tree

(d) *Decision Tree*: Decision trees plays a vital role in the field of medical diagnosis to diagnose the problem [10]. It is a construction of tree structure for designing the decision rules by using group of attributes basing on a selected attribute. It is intended for classification purpose where node represent attributes and branch represent the values assigned for the attribute. Figure 7 show sample Decision Tree.

The advantages and disadvantages of the algorithms stated above are listed in the following Table 1.

## 4 Key Challenges in Handling the Data

The key challenges lies in handling the data and finding the “right” model are as follows.

1. Varying shapes and sizes of data need to be processed.
2. Selecting the right tools for pre-processing the data.
3. Identifying the need of ML for an application.
4. Adapting the right models by experimenting in a Trial and error fashion.
5. For obtaining accurate predictions and results, set of limitations need to be considered and understood.

Finally, a statement can be made that machine learning algorithms are not bound to a specified learning task, rather it is focussed on effective learning. The effective prediction algorithm’s performance is a major challenge which can be resolved by making machine to learn itself with set of classes given as a training set to generate accurate predictions.

**Table 1** Pros and cons of machine learning algorithms

S. No.	Algorithm	Advantages	Disadvantages
1	KNN	<ul style="list-style-type: none"> <li>• Economic Model building</li> <li>• Classification scheme is flexible</li> <li>• Support for Multi-modal classes</li> <li>• Records with multiple class labels are generated</li> </ul>	<ul style="list-style-type: none"> <li>• Unknown records classification is expensive</li> <li>• Does not support on generalization</li> <li>• Thorough computation is required to process the noisy/irrelevant features</li> </ul>
2	Naïve Bayes	<ul style="list-style-type: none"> <li>• Easy implementation</li> <li>• Efficient performance</li> <li>• Less training data</li> <li>• Scales linearly</li> <li>• Continuous and discrete data is handled along with binary and multi-class classification problems</li> <li>• Performs probabilistic predictions</li> <li>• Not sensitive to irrelevant features</li> </ul>	<ul style="list-style-type: none"> <li>• Models need to be trained and tuned properly</li> <li>• Direct application of Naive Bayes is difficult for the “continuous dataset” (like time)</li> <li>• Scaling is not possible when the classes are more</li> <li>• Requires more runtime memory for efficient prediction</li> </ul>
3	SVM	<ul style="list-style-type: none"> <li>• Process semi structured and structured data and can also handle complex function with appropriate kernel function</li> <li>• Less probability of over fitting</li> <li>• High dimensional data scaling is possible</li> </ul>	<ul style="list-style-type: none"> <li>• Performance degrades with respect to training time when large data set is considered</li> <li>• Selection of kernel function is difficult</li> <li>• Not suitable for noisy dataset</li> <li>• Not capable of providing probability estimates</li> </ul>
4	Decision tree	<ul style="list-style-type: none"> <li>• Supports regression, classification problems</li> <li>• Easy interpretation and can handle of categorical and quantitative values</li> <li>• Performs missing values filling in attributes with the most probable value</li> <li>• Ensure high performance due to tree traversal algorithm efficiency</li> </ul>	<ul style="list-style-type: none"> <li>• Unstable</li> <li>• Controlling size of a tree is difficult</li> <li>• Easily Prone to sampling error and generates locally optimal solution a non-global optimal solution</li> </ul>

## 5 Applicability of Machine Learning in Various Fields

The ML algorithms are effective in identifying complex patterns in data of various industries like Finance, Government, Marketing, Transport, and Medical applications. In this research work, the contribution of machine learning strategies with respect to medical applications is performed. In medical applications, Machine learning strategies were being used for several disease diagnosis and detections. Machine learning can deliver alternative decisions regarding the treatment plans for patients.

Several applications include tumour detection in ultra-sonograms, classification and detection of micro calcifications in mammograms, classification of chest X-rays, tissue and vessel classification in Magnetic Resonance Images [11]. In recent times, several research experiments were carried out for the diagnosis and prediction on various diseases like cardiovascular diseases, cancers, and diabetes using various machine learning predictions techniques.

**A. Predictions on Cardio Vascular Diseases** Cardiovascular disease (CVD) is a category of heart diseases which involve the blood vessels. Cardiovascular disease may be CAD like angina and myocardial infarction (heart attack), hypertensive heart disease, stroke, rheumatic, atrial fibrillation, cardiomyopathy, endocarditis, congenital heart disease, aortic aneurysms, venous thrombosis and peripheral artery disease. The key challenge in heart disease is Diagnosis and prediction of heart diseases [12]. The health care research institutes are being enormously working on the aspect of predicting the heart diseases.

The two algorithms namely genetic and neural network algorithms were considered as effective techniques for the prediction of heart disease with the features like family history, age, hypertension, diabetes, cholesterol, alcohol intake, smoking, obesity [13]. The set of methods like Decision tree, Naive Bayes, Neural network algorithms are used for the analysis of medical data sets related to heart disease prediction [14]. Using these three algorithms, a model named Intelligent Heart Disease Prediction System (IHDPS) has been developed. The Results exhibited that each technique has outperformed in realizing the defined objectives [15].

**B. Diabetes Predictions** The major public health issue of all ages in recent years are Diabetes. It is effected when the high blood sugar levels are in rise over a long period. In recent research, it has been identified as a highest risk factor for causing Alzheimer, blindness (Diabetic Retinopathy) and kidney failures. Many researchers made an attempt to develop techniques to identify the causes of diabetes (and or dependent diseases) and treat it (and or them).

The research on diabetes prediction is done by establishing a relationship between diabetes risks using person's daily lifestyle activities like food habits, physical activity, sleeping time, along with other indicators like Body Mass Index (BMI), and waist circumference [16]. The performance analysis of SVM Classifier, Naïve Bayes classifier and Radial Basis Function (RBF) network with respect to the heart, cancer and diabetes datasets expresses that the SVM classifier outperforms the classification [17]. The experimentation is carried out in WEKA environment and the results were observed that SVM acts as a more robust and effective classifier for medical data sets.

**C. Cancer Predictions Using Machine Learning** Cancer is the most universal representation disease where about 100 diseases with various kinds and categories, cancer initially develop as an uncountable and abnormal growth of cells. Several



research over past years is still performing a continuous evolution concerning the prediction and prognosis of cancers like liver, skin, lung and stomach cancers [18, 19]. The set of machine learning and data mining techniques were adopted for cancers prediction and some experimentation were performed using these algorithms on the dataset.

In recent years, the development of optimized models using risk prediction algorithms and techniques were done for the diagnosing breast cancer and recurrence prediction [20, 21]. Several research reviews were made on cancer prognosis prediction along with types and subtypes [22] and the review also suggests the appropriate data sets usage for a cancer type/subtype [23]. A comparative study of machine learning algorithms were performed and among them it is been observed that Naive Bayes classifier is selected as a best model for prognosis of cancer survivability basing on the survival historical data. The Artificial Neural Network (ANN) exhibited best performance in prognosis of breast cancer recurrence [24].

The several machine learning techniques like DT, SVM and ANN are explored to implement the predictive models for recurrence prediction in breast cancer. The performance of the algorithms on the data is assessed through specificity, sensitivity, and accuracy. Most of the researches explored that SVM classification model well predicted breast cancer recurrence with minimal error rate achieving highest accuracy. Here the usage of chosen algorithms i.e., KNN, Naïve Bayes, SVM and DT in the diagnosis of various diseases is presented in Table 2.

## 6 Conclusion

Though several set of Machine Learning algorithms are available, still lot of research and development is going on them. The gap does still exist in selection of model and algorithm basing on the application domain. This paper presents a brief overview of machine learning algorithms and their classification or hierarchy with respect to health care applications. These algorithms when integrated with branch of the AI offers better performance. The performance of each algorithm varies on different platforms. Furthermore, the engineers need to decide the selection of algorithm implementation by considering a number of parameters related to data-sets. The comparative analysis of algorithms need to be made in terms of performance in order to select the best model for the prediction.

**Table 2** Usage of chosen ML techniques in diagnosis of various diseases

Algorithm	Dataset	Accuracy (%)	References
SVM	Pima Indian diabetes	78	Kumari and Chitra [25]
SVM	UCI	98.62	Nazari Kousarrizi et al.[26]
SVM	UCI	99.63	Tyagi et al. [27]
SVM	Wisconsin breast cancer	96.99	Lavanya et al. [28]
NB	Diabetic research institute	86.41	Vembandasamy [29]
NB	Different sectors of society	95	Sarwar and Sharma [30]
NB	Wisconsin	92.42	Shajahaan et al. [31]
NB	Cleveland	85	Yadav et al. [32]
NB	Al-Islam hospital	95.24	Salmi1 et al. [33]
DT	SVNDC hospital	97	Venkatesan et al. [34]
DT +SVM	Wisconsin breast cancer	91	Sivakami et al. [35]
DT	Medical centre Chittagong	73.5	Faisal Faruque et al. [36]
DT	Wisconsin breast cancer	99	Sathiyarayanan et al. [37]
DT	MED-NODE	82.35	Shalu et al. [38]
KNN+RF	Wisconsin breast cancer	100	Rama Devi et al. [39]
KNN	Cleveland UCI	87.1	Palacios et al. [40]
KNN	Wisconsin breast cancer	97.51	Amrane et al. [41]
KNN	Cleveland UCI	85	Pawlovsky et al. [42]

## References

1. Manjunath, The 15 Algorithms Machine Learning Engineers Need to Know. <https://favouriteblog.com/15-algorithms-machine-learning-engineers/>
2. M. Welling, *A First Encounter with Machine Learning* (Donald Bren School of Information and Computer Science, University of California Irvine, 2011)
3. M. Bowles, *Machine Learning in Python: Essential Techniques for Predictive Analytics* (Wiley, Hoboken, NJ, 2018). ISBN: 978-1-118-96174-2
4. S.B. Kotsiantis, Supervised machine learning: a review of classification techniques. *Informatica* 249–268 (2007)
5. D. Nguyen, C. Nguyen, T. Duong-Ba, H. Nguyen, A. Nguyen, T. Tran, Joint network coding and machine learning for error-prone wireless broadcast, pp. 1–7 (2017). <https://doi.org/10.1109/ccwc.2017.7868415>
6. V.S. Kompallia, K.U. Rani. Clusters of genetic-based attributes selection of cancer data, in *Joint Network Coding and Machine Learning for Error-Prone Wireless Broadcast*, pp. 1–7 (2017). <https://doi.org/10.1109/ccwc.2017.7868415>
7. Unsupervised ML. <https://www.onclick360.com/unsupervised-machine-learning/>
8. Extending ML Algorithms—KNN. <https://www.youtube.com/watch?v=NpwjJ28up28>
9. SVM—Soft Margin Formulation and Kernel Trick. <https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe>
10. D. Lavanya, K. Rani, Performance evaluation of decision tree classifiers on medical datasets. *Int. J. Comput. Appl.* **26** (2011). <https://doi.org/10.5120/3095-4247>
11. K. Usha Rani, Analysis of heart diseases dataset using neural network approach. *Int. J. Data Min. Knowl. Manag. Process (IJDKP)* **1**(5) (2011)

12. S. Mendis, P. Puska, B. Norrving, WHO, Global atlas on cardiovascular disease prevention and control, pp. 3–18 (2011) ISBN 978-92-4-156437-3
13. S.U. Amin, K. Agarwal, R. Beg, Genetic neural network based data mining in prediction of heart disease using risk factors, in *IEEE Conference on Information & Communication Technologies (ICT)*, pp. 1227–31. 11–12 Apr 2013
14. M. Gandhi, S.N. Singh, Predictions in heart disease using techniques of data mining, in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pp. 520–525. 25–27 Feb 2015
15. S. Palaniappan, R. Awang, Intelligent heart disease prediction system using data mining techniques, in *IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2008*, pp. 108–115. 31 Mar–4 Apr 2008
16. A. Anand, D. Shakti, Prediction of diabetes based on personal lifestyle indicators, in *1st International Conference on Next Generation Computing Technologies (NGCT)*, pp. 673–676. 4–5 Sept 2015
17. P. Janardhanan, L. Heena, S. Fathima, Effectiveness of support vector machines in medical data mining. *J. Commun. Soft. Syst.* **11**(1), 25–30 (2015)
18. A. Akutekwe, H. Seker, S. Iliya, An optimized hybrid dynamic Bayesian network approach using differential evolution algorithm for the diagnosis of Hepatocellular Carcinoma, in *2014 IEEE 6th International Conference Adaptive Science & Technology (ICAST)* (IEEE, 2014)
19. K. Kourou et al., Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015)
20. C.A. Drukker, Optimized outcome prediction in breast cancer by combining the 70-gene signature with clinical risk prediction algorithms. *Breast Cancer Res. Treat.* **145**(3), 697–705 (2014)
21. A. Bhardwaj, A. Tiwari, Breast cancer diagnosis using genetically optimized neural network model. *Expert Syst. Appl.* **42**(10), 4611–4620 (2015)
22. J. Das, K.M. Gayvert, H. Yu, Predicting cancer prognosis using functional genomics data sets. *Cancer Inf.* **13**(Suppl), 5 (2014)
23. B.R. Cirkovic et al., Prediction models for estimation of survival rate and relapse for breast cancer patients, in *IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 1–6. 2–4 Nov (2015)
24. A.T. Eshlaghy et al., Using three machine learning techniques for predicting breast cancer recurrence. *J. Health Med. Inf.* (2013)
25. V.A. Kumari, R. Chitra, Classification of diabetes disease using SVM. *Int. J. Eng. Res. Appl.* **3**, 1797–1801 (2013)
26. M.R. Nazari Kousarrizi et al., An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification. *Int. J. Electr. Comput. Sci.* **1**, 13–19 (2012)
27. A. Tyagi et al., in *5th IEEE International Conference on Parallel Distributed and Grid Computing (PDGC-2018)* (Solun, India, 2018), pp. 689–693
28. D. Lavanya et al., Analysis of feature selection with classification: breast cancer datasets. *Int. J. Comput. Sci. Eng. (IJCSE)* **2**, 756–763 (2011)
29. K. Vembandasamy et al., Heart Diseases Detection Using NB Algorithm, vol. 2 (2015)
30. A. Sarwar, V. Sharma, Intelligent Naive Bayes approach to diagnose diabetes type-2. *IJCA Spec. Issue Issues Challenges Netw. Intell. Comput. Technol. ICNICT* **3**, 14–16 (2012)
31. S. Shajahaan et al., Application of data mining techniques to model breast cancer data. *Int. J. Emerg. Technol. Adv. Eng.* **3** (2013)
32. A. Yadav et al., Better healthcare using machine learning. *Int. J. Adv. Res. Comput. Sci.* **9**(3) (2018)
33. N. Salmi1 et al., Naïve Bayes classifier models for predicting the colon cancer. *IOP Conf. Ser. Mater. Sci. Eng.* (2019)
34. E. Venkatesan et al., Performance analysis of decision tree algorithms for breast cancer classification. *Indian J. Sci. Technol.* **8**, 1–8 (2015)
35. K. Sivakami et al., Mining big data: breast cancer prediction using DT, SVM hybrid model. *Int. J. Sci. Eng. Appl.Sci.* **1** (2015)

36. M.D. Faisal Faruque et al., Performance analysis of machine learning techniques to predict diabetes mellitus, in *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 7–9 Feb 2019
37. P. Sathiyarayanan et al., Identification of breast cancer using the DT algorithm, in *Proceeding of International Conference on Systems Computation Automation (2019)*
38. Shalu et al., A color-based approach for melanoma skin cancer detection, in *First International Conference on Secure Cyber Computing and Communication (ICSCCC)* (2018)
39. G.N. RamaDevi, K.U. Rani, D. Lavanya, in *Ensemble-Based Hybrid Approach for Breast Cancer Data*, ed by A. Kumar, S. Mozar. ICCCE 2018. Lecture Notes in Electrical Engineering, vol. 500 (Springer, Singapore, 2019)
40. A.P. Palacios et al., An immune algorithm that uses a master cell for component selection for the KNN method, in *Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)* (2019)
41. M. Amrane et al., Breast cancer classification using machine learning, in *IEEE* (2018)
42. A.P. Pawlovsky et al., An ensemble based on distances for a kNN method for heart disease diagnosis

# A Methodology for Detecting ASD from Facial Images Efficiently Using Artificial Neural Networks



T. Lakshmi Praveena and N. V. Muthu Lakshmi

**Abstract** Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder. Neurodevelopmental disorders are related to the brain development and consequent changes occur in facial tissues. The changes in facial tissues lead to changes in face landmarks. Facial landmarks are the pin points in face that helps to identify different parts in face. ASD individuals have differences in facial landmarks compared to non ASD individuals of similar age group due to the developmental delay in brain. Effective and reliable algorithms to process facial images are artificial neural networks (ANN). Dataset for present research study are collected from autism parenting group and from other web sources. Collected dataset includes male and female ASD and non ASD individuals of 1–10 years of age. Present research helps parents, pediatricians, neurologists to assess and detect ASD in kids and also to analyze ASD severity in individuals. The early detection and analysis of ASD helps to give better treatment and to give better life to ASD children.

**Keywords** Autism spectrum disorder · Facial land marks · Artificial neural networks · Machine learning

## 1 Introduction

Autism Spectrum Disorder (ASD) is a neurological development disorder. Neurological development is inter-related with brain development. Brain development is affected at the time of prenatal period and brain development can also be affected by the postnatal conditions of kid. The parental age, parental weight, smoking, alcohol consumption at prenatal stage affects brain development and it can lead to ASD [1]. The postnatal conditions like low birth weight, head weight and head circumference

---

T. Lakshmi Praveena (✉) · N. V. Muthu Lakshmi  
Sri Padmavathi Mahila Visvavidyalayam, Tirupati, AP, India  
e-mail: [praveenalaxmi1@gmail.com](mailto:praveenalaxmi1@gmail.com)

N. V. Muthu Lakshmi  
e-mail: [nvmuthulakshmi@gmail.com](mailto:nvmuthulakshmi@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_31](https://doi.org/10.1007/978-3-030-46939-9_31)

differences can effect brain development and lead to ASD in future developmental age [2]. The development of brain can be observed in facial landmarks of ASD individuals compared to non ASD individuals. ASD individuals have differences in facial landmarks points and coordinates [3]. The changes of brain are reflected in face, such as “Face predicts the brain” [4]. Finding one or more facial landmarks differences in ASD individuals helps as biomarker in diagnosis of ASD. Facial landmarks are different for various ASD and developmental disability individuals. As initial step of research, facial images are preprocessed to remove noise in images. In second step facial landmarks are identified by using neural networks algorithms. The retrieved landmarks and their coordinates are created as matrix. Coordinates matrix is analyzed with help of Euclidian distance measure. The performances of generated results are analyzed using statistical measures. The further sections of paper are arranged as; Sect. 2 gives overview of ASD, ANN algorithms and methods. Section 3 of paper has overview of proposed method and procedure. Section 4 of paper provides implementation and result analysis.

## 2 Literature Review

### 2.1 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a neurodevelopmental disorder. In general, neurodevelopmental disorders symptoms are hyperactivity, deficiency in social communication, deficiency in learning and language. America’s children and the environment (ACE) published an article on neurodevelopmental disorders and also detailed review of ASD and other disorders [5]. Diagnosis process is manual process conducted by multiple streams of doctors by observing the behaviour of a child. Actually, ASD is detected at the age of 3 years or above 3 years [6] with manual diagnosis. Treatments started at this level takes long time to show benefits. For the past few years, many research works have done to speed up the diagnosis process so that ASD detection can be done at an early age which improves the curing mechanism. The standardized diagnosis process of ASD includes assessment of Intellectual Quotient (IQ), Verbal IQ using Childhood Autism Rating Scale (CARS), Autism Diagnosis Instrument—Rating (ADI-R) and Weschler Abbreviated Intelligence Scale (WAIS) diagnosis tools. As a part of clinical diagnosis brain MRI, EEG and physical examination of face. Researches of ASD states that ASD is result of embryological brain development deficiency which is reflected in development of physique or face in ASD individuals [7]. The main objective of present research is to detect the facial biomarkers in ASD individuals, also to find correlation between facial anomalies and ASD symptoms in individuals. Researchers have worked with many techniques to address this issue among many; machine learning is one, which is efficient and reliable to detect ASD with less processing time. Machine learning can train the system with past data and then machine will predict ASD within short period of time.

## ***2.2 Artificial Neural Networks in ASD Detection***

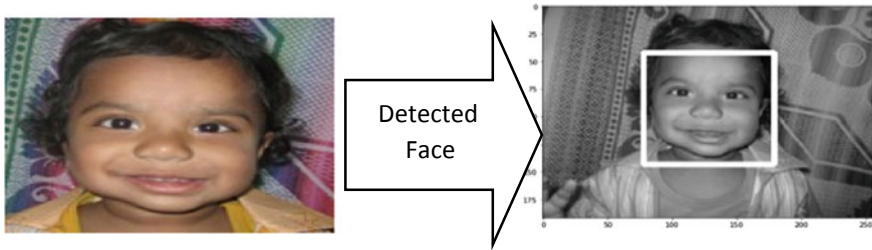
ANN is a network of interconnected neurons similar to brain network. Neurons are artificial neurons which act as processing units, this artificial neuron learns by adjusting its weights [8]. ANN algorithms are used to adjust artificial neurons for better learning and accuracy. ASD is a neurodevelopment disorder, connected with the brain activity. The anomalies in brain cause ASD. The brain architecture and working is similar to neural networks. Artificial neural network is one of the types of neural networks [9–11]. Lamyaa Sadouk [12] proposed a novel deep learning approach for analyzing ASD individuals' motor movements using Convolution Neural Networks (CNN) based classification algorithm. AL-Allaf [13] has reviewed on ANN algorithms for face detection and has compared different algorithms performance. Next section of paper discusses about facial landmarks detection and detecting ASD with facial landmarks using ANN algorithms.

## **3 Proposed Methodology**

ASD individual has different facial landmark patterns, different distance between landmarks and has difference width of landmarks compared to typical developed (TD) people [7]. By using advanced imaging technology and neural networks, facial landmarks are detected and analyzed based on proposed criteria by ASD researchers [7, 13]. The problem statement is to build a detection Methodology for ASD by analyzing facial landmarks of ASD individuals using ANN algorithms. The proposed methodology is divided into different phases like face detection from images, face landmark detection, transforming landmarks as features, feature extraction, applying ANN on facial landmark data and evaluation of performance by comparing with other machine learning algorithms.

### ***3.1 Dataset Collection and Preprocessing***

Proposed methodology is using 50 ASD images collected from Autism Parenting Hub face book group and 50 images of TD group are collected from online sources of ages from 1 to 10 years. Collected images are preprocessed using image processing algorithms to remove noise, resize image to required dimension. Images are initially converted to gray scale images. Gray scale images are divided into frames to detect face region based on pixel intensity. Preprocessing of images is done by using image processing algorithms supported by OpenCV library.



**Fig. 1** Face detection from input image

### 3.2 Detecting Face Region from Image

Face detection is the initial step in processing faces for different applications. Face detection is useful in biometrics applications, security checking applications, emotion recognition applications, artificial intelligence applications [13]. Face detection algorithm is initially proposed by viola and jones [14], in this research adaboost classifier based with haar features are designed for face detection and this classifier also provides different objects haar features for object detection. Implementation of face detection is performed by OpenCV and Dlib library using python language. ASD and TD faces of dataset are given as input to face detection algorithm to find face region as shown in Fig. 1.

#### 3.2.1 Face Landmarks Detection from Extracted Face Region

Facial landmarks detection has applications in emotion analysis, face identity, facial animation. Facial landmarks detection algorithms are used to identify key points in face and compare or classify faces. ANN algorithms are used in facial landmark detection. Frequently used algorithms are Convolution Neural Networks (CNN), Restricted Boltzmann Machine (RBM), Auto encoders [15]. RBM provides face predictor with 68 face landmarks and are stored as vector of points. Face landmarks coordinates are differed by head pose and expression. Figure 2 shows the 68 landmarks and their position in face. Detected landmarks for ASD individual using the model implemented is shown Fig. 3.

### 3.3 Feature Extraction of from Face Landmarks

Feature extraction from face image is another important phase in analyzing images. Feature extraction in standard image processing uses different filters as feature extraction algorithms like Local Binary Pattern (LBP), Gabor Filters [16]. Proposed methodology extracts 8 important features by calculating mean value of landmark





Fig. 2 68 facial landmarks identified by RBM predictor

Fig. 3 Landmarks detected in ASD&TD



points of specific region and extracts 15 important features by finding Euclidean distance between landmarks points of required features. Extracted features are listed in Tables 1 and 2. The important and correlated to ASD features are selected for classification and detection of ASD.

### 3.4 Statistical Analysis of Dataset

Structured dataset is constructed by transforming the landmarks using statistical measures applied on landmarks points. Mean is calculated to transform and construct 8 feature dataset. Euclidean distance is calculated to transform and construct 15

**Table 1** 8 feature face landmarks dataset statistics

Feature No	Facial landmark	Facial landmark points	ASD (Mean $\pm$ SD)	TD (Mean $\pm$ SD)
1	Outer mouth	48–59	274.2 $\pm$ 55.5	281.5 $\pm$ 56.7
2	Inner mouth	60–67	273.5 $\pm$ 55.6	280.8 $\pm$ 56.5
3	Right eyebrow	17–21	192.7 $\pm$ 51.6	192.4 $\pm$ 31.6
4	Left eyebrow	22–26	236.2 $\pm$ 51.7	239.7 $\pm$ 43.9
5	Right eye	36–41	210.2 $\pm$ 50.9	210.3 $\pm$ 35.3
6	Left eye	42–47	252.6 $\pm$ 54.4	255.5 $\pm$ 48.3
7	Nose	27–34	248.6 $\pm$ 52.6	254.7 $\pm$ 48.7
8	Jaw	0–16	269.7 $\pm$ 55.4	275.3 $\pm$ 52.8

**Table 2** 15 feature face landmarks dataset statistics

Feature No	Facial landmark	Facial landmark width	ASD (Mean $\pm$ SD)	TD (Mean $\pm$ SD)
1	Fore head width	Dist(0,16)	178 $\pm$ 68	189 $\pm$ 56
2	Eye outer width	Dist(45,36)	144 $\pm$ 65	142 $\pm$ 51
3	Eye inner width	Dist(42,39)	115 $\pm$ 72	99 $\pm$ 60
4	Left eye width	Dist(39,36)	107 $\pm$ 74	108 $\pm$ 66
5	Right eye width	Dist(45,42)	138 $\pm$ 86	121 $\pm$ 74
6	Right face width	Dist(45,27)	136 $\pm$ 79	120 $\pm$ 67
7	Left face width	Dist(36,27)	136 $\pm$ 79	120 $\pm$ 67
8	Nose width	Dist(35,31)	142 $\pm$ 90	109 $\pm$ 67
9	Mouth width	Dist(54,48)	119 $\pm$ 80	143 $\pm$ 79
10	Nose height	Dist(33,27)	120 $\pm$ 71	104 $\pm$ 56
11	Cheek height left	Dist(48,36)	140 $\pm$ 91	147 $\pm$ 88
12	Cheek height right	Dist(54,45)	144 $\pm$ 68	131 $\pm$ 55
13	Upper lip dist	Dist(63,33)	126 $\pm$ 88	117 $\pm$ 77
14	Upper lip height	Dist(62,51)	128 $\pm$ 95	123 $\pm$ 84
15	Lower lip height	Dist(66,57)	138 $\pm$ 103	133 $\pm$ 89

feature dataset. The same process is applied for ASD and TD images. Sample dataset statistics are given in Tables 1 and 2. Statistical analysis Table 1 shows the mean and standard deviation (SD) of 8 features. The mean value represents the normal and  $\pm$ SD value represents subnormal and abnormal conditions of ASD. ASD features mean value is  $\leq$  to the mean value of TD mean value in all 8 face landmarks. The face features outer mouth, inner mouth, jaw and nose has difference in mean value for ASD and TD. Face features right eye brow and right eye, left eye, left eye brow are equal and has less difference in mean value between ASD and TD. Table 2 shows

the mean and SD values of 15 facial features. Mean values of forehead width, nose width and left cheek height are < in ASD than TD values. These features plays main role in detecting ASD.

### 3.5 Applying ANN on Extracted Feature Dataset

ANN algorithms are used to build classifier model to detect ASD from extracted feature dataset. Proposed methodology uses multilayer ANN and one input layer, two dense hidden layers, one output layers are used. Random weights are generated for first hidden layer and added with negligible bias. Hidden layers are constructed with 4 neurons in each layer. Different activation functions are applied for different layers. Activation functions in ANN are used to find weighted sum of input data vector, weight matrix and bias value. Activation function decides when the neuron to be fired. Activation functions are either linear or non linear functions [17, 18]. Prediction accuracy of Artificial Neural Networks by using different activation functions in different layers is given in Table 3. Proposed methodology is implemented with tensorflow and keras library using python language. Proposed methodology creates dense layers with 4 neurons in hidden layers each and one neuron in output layer. The output layer value ranges from 0 to 1, less than equal to 0.5 is considered as TD i.e. classified as NOASD and greater than 0.5 is considered as ASD. Table 1 presents prediction accuracy of 8-features dataset and 15 features dataset of 50 ASD individuals, 50 TD individuals. Prediction accuracy is improved with the combination of activation functions in different layers. The best combination of activation functions are, first best combination is Selu in hidden layer 1, Relu in hidden layer 2 and sigmoid in output layer with 63% accuracy for 8-features dataset and 75% with 15-features dataset. Second best combination is Relu in hidden layer, Selu in hidden layer 2 and sigmoid in output layer with 70% accuracy for 8-features dataset and 70% accuracy with 15 features dataset. More than 70% of test data is successfully detected as ASD with ANN and facial landmarks. The main landmarks effected in

**Table 3** Prediction accuracy of ANN using different activation functions in different layers

S. No	Hidden layer 1 (activation function)	Hidden layer 2 (activation function)	Output layer (activation function)	Accuracy of ANN on 8 features dataset (%)	Accuracy of ANN on 15 features dataset (%)
1	Relu	Relu	Sigmoid	63	62
2	<b>Selu</b>	<b>Relu</b>	<b>Sigmoid</b>	<b>63</b>	<b>75</b>
3	<b>Relu</b>	<b>Selu</b>	<b>Sigmoid</b>	<b>70</b>	<b>70</b>
4	Softmax	Selu	Sigmoid	70	56
5	Linear	Softmax	Sigmoid	70	70

Bold indicates that these are the best activation functions which are giving better accuracy than other methods

**Table 4** Comparison of proposed methodology with other machine learning algorithms

	Accuracy (%)	MSE	Rsquare
<i>Face_8_features</i>			
Linear regression	31	0.29	-0.16
Logistic regression	53	0.47	-0.88
Decision trees	53	0.47	-0.88
K-nearest neighbour	71	0.15	-0.1
Artificial neural networks	81	0.15	-0.1
<i>Face_15_features</i>			
Linear regression	35	0.25	-0.15
Logistic regression	55	0.42	-0.8
Decision trees	58	0.42	-0.88
K-nearest neighbour	74	0.14	-0.1
Artificial neural networks	82	0.15	-0.1

detection of ASD are mouth width, difference between eyes, eye brows, larger jaw, left cheek height, outer eye distance, nose height and length. These differences show the level of ASD in individuals, and are less in less than 3 years, more in above 3 years age.

### 3.6 Comparative Analysis

Table 4 gives comparative analysis of popular machine learning algorithms with ANN. Facial dataset is used build classifier to predict ASD from facial features. Compared to other algorithms ANN gives more than 80% accuracy with less mean square error (MSE) and R square error. The hypothesis of present methodology is proved to be true and satisfied. Facial landmarks and facial features are biomarker to detect ASD in children at the age of <5 years with maximum accuracy.

## 4 Conclusion

An artificial neural network for diagnosing autism was proposed. The input factors were obtained from an autism data set with face images of ASD and TD. The model was tested and the total result was more than 80% accuracy. This study showed the ability of the artificial neural network to diagnose ASD using facial landmarks of individuals with less computational time and better accuracy. ANN plays an important role in prediction, detection and classification of data. ANN gives very good

results in disease prediction and detection in health care. The proposed methodology can be used on age groups of >10 years. Present methodology proposes a biomarker for detection ASD. It can also be applicable to other neurodevelopmental disorders. Further the work can be extended to other biomarkers of ASD diagnosis like MRI, EEG, and speech.

## References

1. A. Mezzacappa, P.A. Lasica, F. Gianfagna, O. Cazas, P. Hardy, B. Falissard, A.L. Sutter-Dallay, F. Gressier, Risk for autism spectrum disorders according to period of prenatal antidepressant exposure a systematic review and meta-analysis, *JAMA Pediatr.* **171**(6), 555–563 (2017). <https://doi.org/10.1001/jamapediatrics.2017.0124> Published online 17 Apr 2017
2. C. Wang, H. Geng, W. Liu, G. Zhang, Perinatal, perinatal, and postnatal factors associated with autism a meta-analysis, *Medicine* **96**(18), e6696 (2017). <http://dx.doi.org/10.1097/MD.0000000000006696>
3. K. Aldridge, I.D. George, K.K. Cole, J.R. Austin, T.N. Takahashi, Y. Duan, J.H. Miles, Facial phenotypes in subgroups of prepubertal boys with autism spectrum disorders are correlated with clinical phenotypes. *Mol. Autism* **2**, 15 (2011)
4. W. DeMyer, W. Zeman, C.G. Palmer, The face predicts the brain: diagnostic significance of median facial anomalies for holoprosencephaly (arhinencephaly). *Pediatrics* **34**, 256–263 (1964)
5. F. Thabtah, D. Peebles, A new machine learning model based on induction of rules for autism detection. *Health Inf. J.* 1–23. <https://doi.org/10.1177/1460458218824711>
6. D. Bone, M.S. Goodwin, M.P. Black et al., Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *J. Autism Dev. Disord.* **45**(5), 1–16 (2014)
7. G. Tripi et al., Cranio-facial characteristics in children with autism spectrum disorders (ASD). *J. Clin. Med.* **8**(5), 641 (2019). <https://doi.org/10.3390/jcm8050641>
8. J.S. Norris. Face detection and recognition in office environments. Department of Electrical Engineering and Computer Science thesis, Massachusetts Institute of Technology
9. P.N. Belhumeur et al., Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, pp. 713–714 (1997)
10. C. Ding, D. Tao, A comprehensive survey on pose invariant face recognition. Available: <https://arxiv.org/abs/1502.04383>
11. L. Wiskott, J.-M. Fellous, N. Kuiger, C. von der Malsburg, Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 775–779 (1997)
12. O.N.A. Al-Allaf, Review of face detection systems based artificial neural networks algorithms. *Int. J. Multimedia Appl. (Ijma)* **6**(1) (2014)
13. Z. Lei, S.Z. Li, in, Face recognition models: computational approaches. *Int. Encycl. Soc. Behav. Sci.* (2015)
14. P. Viola, M. Jones, Face recognition by humans, in *Face Processing* (2006)
15. M. Bodini, A review of facial landmark extraction in 2D images and videos using deep learning. *Big Data Cogn. comput.* **3**, 14 (2019). <https://doi.org/10.3390/bdcc3010014>
16. D. David-Vico, Deep neural networks. Master's thesis, Autonomous University of Madrid
17. T.H. Le, Applying artificial neural networks for face recognition. *Adv. Artif. Neural Syst.* **2011**, (673016), 16 (2011). <https://doi.org/10.1155/2011/673016>
18. M. Kirby, L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Tran. Pattern Anal. Mach. Intell.* **12**(1) (1990)

# Service Composition in Mobile Ad Hoc Networks (MANET's) with the Help of Optimal QoS Constraints



G. Manoranjan, M. V. Rathnamma, V. Venkata Ramana, and G. R. Anil

**Abstract** In recent year's usage of mobiles, laptops, PDA's etc., are greatly increased. Usage of implementation SOA (service oriented architecture) is to increase flexibility in providing the services in MANET's (Mobile Ad hoc networks). Due to dynamic topology, resource heterogeneity, bandwidth, providing a service provider is a critical challenge. Existing composition services are not suitable for the usage of constraints consideration while choosing a service in the proposed system we uses Response time, throughput, energy consumption, hop count as a QoS constraints to provide the respective services. Based on fuzzy logic we are evaluating rating of a service according to QoS service constraints. From rating we can able to provide the services. Nodes in the MANET's can provide more than one service. In the simulation results we used to compare our proposed method with the AODV protocol and we got better performance results in respective of average packet delay, Energy constraint, Throughput and Turnaround time.

**Keywords** Fuzzification · Defuzzification · Service provider · Service Request-or · MANET's

## 1 Introduction

A Mobile Ad hoc network (MANET) is also known as wireless Ad hoc networks or Ad hoc wireless network. It is a continuously self-configuring infrastructure less network of mobile devices connected with the help of wireless connections. In the MANET's nodes have freedom of move independently in any direction and change its

---

G. Manoranjan · G. R. Anil  
SCIS, University of Hyderabad, Hyderabad, India  
e-mail: [manoranjangandhudi@gmail.com](mailto:manoranjangandhudi@gmail.com)

M. V. Rathnamma (✉)  
Department of CSE, KSRM, Kadapa, India  
e-mail: [rathnamma@ksrmce.ac.in](mailto:rathnamma@ksrmce.ac.in)

V. Venkata Ramana  
Department of CSE, CBIT, Proddatur, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_32](https://doi.org/10.1007/978-3-030-46939-9_32)

links to other networks frequently. In this paper we are considering the two dependent components such as service provider (SP) and service request-or (SR). The MANET's can be available of different types and these can be used in different applications such as vehicular based Ad hoc networks, smart phone based Ad hoc networks, internet based Ad hoc networks, hub spoke MANET's, military or tactical networks and flying Ad hoc networks. In the existing system based on the available SP's the SR can formulate a service plan to the service composition and to achieving the multiple objective (MOO) optimization problem [1, 2]. There is no centralized control for the MOO problem. In terms of space complexity in the service composition and service binding and maintenance also more complicated [3]. This problem can be arrived due to malicious behavior of nodes and received information is erroneous, incomplete and uncertain in the MANET environment. In the multiple objective optimization (MOO) problem they uses the trust based service composition and binding [4, 5]. Because of lack of central authority server such as mobility, wireless communication links providing a security in a MANET is a challenging task. Our work is to propose a dynamic trust based composition and binding algorithm by using the fuzzy based optimal QoS constraint service oriented Mobile Ad hoc Networks. In the MOO problem they uses the integer linear programming (ILP) solution and this solution can have exponential run time complexity and only applicable for the small sized problem and in the existing system.

## 2 Related Work

In the service composition and binding we mainly have two different objectives to provide the service such as

**(i) Goal oriented composition** In the process of goal oriented composition a set of services can be composed in such a manner to achieve the goals. In this paper we have a several set of services requests and to provide the services to those goals. By achieving the goals we need to provides the minimum cost for the services, reliability of service, and to maintain the fault tolerance. Fault tolerance is means that operate the system properly.

**(ii) Work Flow based Composition** Consider the work flow with a given constraints as inputs. Our objective of the work takes the latter approach with the help of service composition as workflow problem Rehman and Kim [6] implements the novel protocol. Which is named as location-aware on-demand multi-path catching and forwarding for ND N-based MANET's? Wu et al. [7] stated that disruption loss, interference, and jamming can be improved significantly by using network coding (NC). Khan et al. [8] suggested that identifying the malicious nodes by using the network parameters with the packet loss in the network. Taha et al. [9] stated that by using the fitness function we can optimize the energy consumption in Ad hoc on demand routing protocol.

### 3 System Model

#### 3.1 SP-SR Model

In this service-oriented Mobile Ad hoc Network we mainly consider two functions to achieve the corresponding goals.

- (I) Service provider (SP) which can provide the services based on available resources and as well as requests. Here service provider works based on user location. Based on changes in user location the service provider can dynamically changes its path based on the request.
- (II) Service request-or can have the capable of sending the request based on behalf of its owner and here based on changes in the location the service request-or can able to request the services. Consider an example such that in a smart city a person wants to visit nearby fancy store after the purchase he wants to move to vegetable market. For this we need to identify the fancy store nearby vegetable market. Here we mainly consider the user location and various fancy stores which are located nearby vegetable market. When the request placed by SR they mainly mentioned on its menu such as QoI, QoS and cost for the service request. (e.g., famous fancy store, fresh vegetable market and less transportation i.e., duration of travel) and individual abstract services (E.g., powder).

#### 3.2 Service Constraints

By considering the execution time, latency, response time, transmission time, W3C performance is to be defined. Service performance can be denotes that how faster the given service can be completed. In this paper we are considering the throughput and response time are the two major constraints.

**Response time** The response time is to be considered such as time duration between the sending of a request to the receiving of a response. in between duration we are mainly considering the service processing time, network processing time, time consumed for compression and decompression, and time consumed for the encryption and decryption of data and time for data traversing through protocol stack of source, intermediate and destination nodes mathematically response time is denoted here.

$$t_{\text{response}}(s) = t_{\text{task}}(s) + t_{\text{stack}}(s) + t_{\text{transport}}(s) + t_{\text{cd}}(s) + t_{\text{ed}}(s) \quad (1)$$

$t_{\text{response}}(s)$  response time or a given service  
 $t_{\text{task}}(s)$  processing time for a task



$t_{\text{stack}}(s)$	time consumed for processing of data in protocol stacks of source, destination and intermediate nodes
$t_{\text{cd}}(s)$	time required for compression and decompression of data
$t_{\text{ed}}(s)$	time required for encryption and decryption

### Throughput

Throughput can be defined as number of requests to be completed in a given time is calculated as throughput.

$$\text{Throughput}(s) = (\#\text{requests})/(\text{time}) \quad (2)$$

### 3.3 Node Qos Constraints

We are considering two parameters to calculate the corresponding performance of a service such as energy and hop count.

#### Energy consumption for a node

The energy consumption can be done when the packets are sending from one node to another node. The energy consumption of a node is equally denoted as each node energy consumption is calculated as

$$E(p, n_a) = E_{tx}(p, n_a) + E_{rx}(p, n_b) + (N + 1)E_0(p, n_i) \quad (3)$$

$$E_{\text{Node}} = E_{\text{ack}} + \sum_{i=1} \text{Cost}_{E_i}$$

$$E_{\text{ack}} = n \times E(p, n_a) \quad (4)$$

where  $n$  is the number of controlled packets and  $E_{\text{ack}}$  is the time consumed for processing of data in protocol stacks of source, destination, and intermediate nodes, cost  $E_i$  can be denoted as the cost incurred for distinct mobility constraints, and  $E_i$  denotes the processing of node movement, resources, service discovery and bandwidth.

#### Hop Count

The number of hops or links to be presented in between the source and destination node can be defined as hop count. Average hop count is considered for overall communicating nodes in MANET's to compute the average shortest path hop count at each point in time. We use the multi-hop connectivity matrix. Hop count  $h$  is

$$h = \frac{\sum_{i=1}^T \text{hops}_i}{\sum_{i=1}^T \text{paths}_i} \quad (5)$$

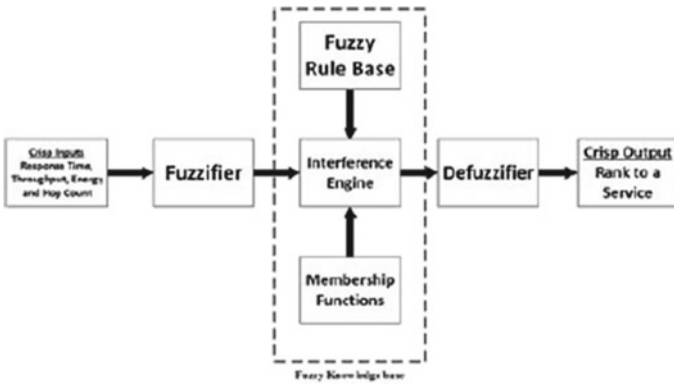


Fig. 1 Flow chart to finding rating to the service

where  $T$  = the number of multi-hop matrix

Hops<sub>*i*</sub> = at time I total number of hops and paths<sub>*i*</sub> = number of cells at time I have non-zero entry.

## 4 Fuzzy Interface System to Calculate the Rating of a Service

To calculate the rating for a service we consider energy, hop-count and service related throughput and response time as a input parameters.

### 4.1 Procedure for Calculate the Rating to the Service

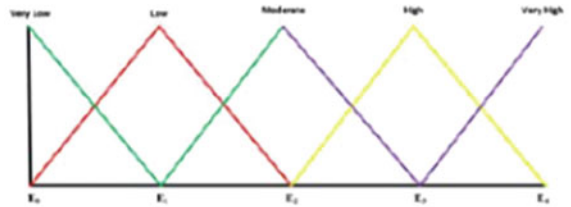
Fuzzification: This methodology which compares the given input values to the corresponding membership function and to formulate the membership values from the past history and them converted into linguistic values. In the proposed methodology we consider the input variables as the energy, hop count, throughput, response time and the output as the rating to the service i.e., fuzzy output variable (Fig. 1).

### 4.2 Fuzzy Membership Function

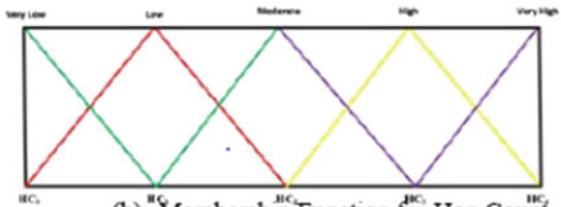
Here, we considered input variables (hop count, throughput, energy, response time) are divided into five different sets such as high, moderate, very high, low, very low.

Then output crisp values can be generated by fuzzy functions and here we are going to consider the triangular membership function to finding the rating for the service (Fig. 2).

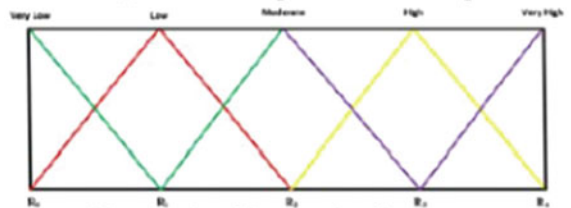
Fig. 2 Fuzzy membership functions



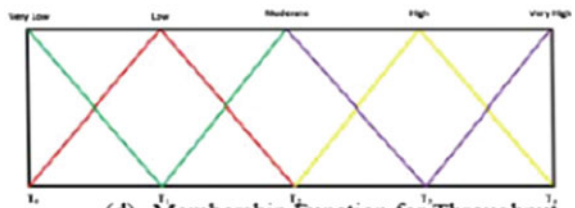
(a) Membership Function for Energy



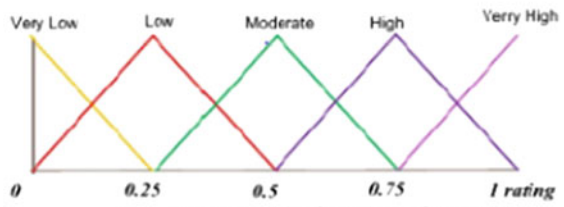
(b) Membership Function for Hop Count



(c) Membership Function for Response Time



(d) Membership Function for Throughput



(e) Membership functions for rating

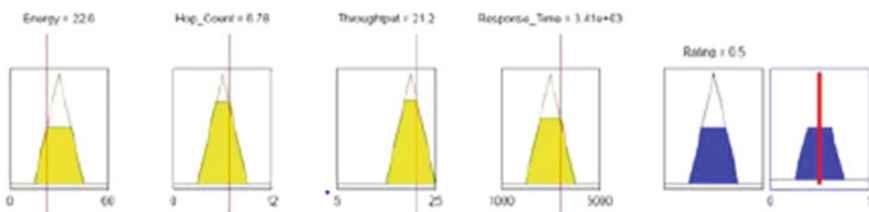
### 4.3 Fuzzy Rule Base

This is considered as a database to find the rating for the service. Here we are consider the combination of fuzzy input sets. Each rule can be associated by “if-then”. And we have 5 fuzzy input variables so fuzzy rule base contains of  $625(5 * 5 * 5 * 5)$  rules. Some of the rules are tabulated in Table 1.

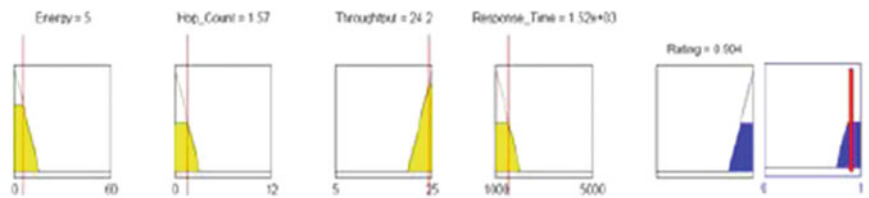
**Evaluating a rating of a service using fuzzy interface system** By using the Fuzzification we use to combine all the crisp inputs such as energy, response time, hop count and throughput. In Fig. 3, Fuzzification can be applied on the two fuzzy rules and fuzzy sets can be explained in Table 2.

**Table 1** Fuzzy rules

Energy	Hop count	Response time	Throughput	Rating
Very low	Very low	Very low	Very high	Very high
High	Low	Moderat	Low	Very high
Low	Low	Low	High	High
Low	Moderat	Low	High	High
Moderat	Low	Moderat	High	Moderat
Moderat	Moderat	Moderat	High	Moderat
High	High	Moderat	Low	Low
Very high	high	high	Low	Very low
Very high	Very high	Very high	Very low	Very low



(a) Rule 1



(b) Rule 2

**Fig. 3** Finding rating for the service with defuzzification

**Table 2** QoS parameters for fuzzy sets

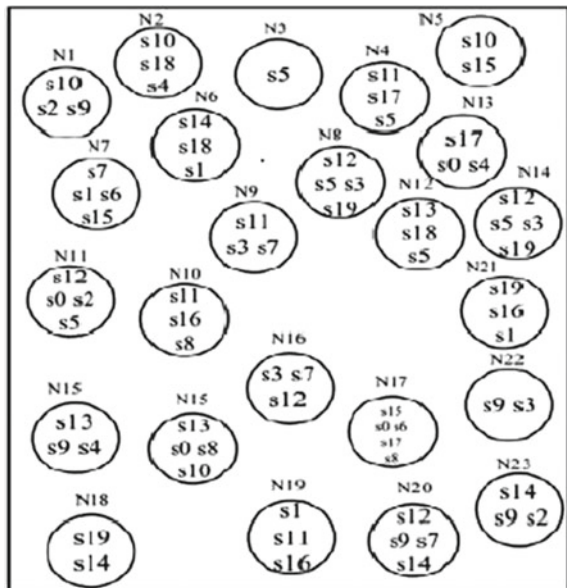
QoS parameters	Very low	Low	Moderate	High	Very high
Energy (J)	0–15	0–30	10–40	20–50	30–60
Hop count	0–3	0–6	3–9	5–12	9–15
Response time (ms)	1000–2000	2000–3000	3000–4000	3000–5000	4000–5000
Throughput	5–10	5–15	10–20	20–25	25–30

**Set of fuzzy rules**

- (i) RULE-1: (throughput is high, energy, hop count, and response time is low) such that rating for a service is high.
- (ii) RULE-2: (throughput is very low, energy, hop count, and response time is very high,) such that rating for a service is very low.

To assign the input crisp values of energy, hop-count, throughput and response time of an application are to like 5 J, 1.57, 24.2 and 1.52e +03 accordingly as shown in Fig. 4a. And find the membership value degree for every given input such as intersecting point if fuzzy triangular wave. Identify the minimal membership degree value of all input values and consider corresponding shaded portion in the output fuzzy set. Union all output fuzzy set’s shaded portion and apply the center of gravity method to evaluate crisp node and service rating. Similarly rule 2 also represented as shown in Fig. 3b.

**Fig. 4** Nodes in the MANET’s with multiple services



#### 4.4 Defuzzification

In this the output of the fuzzy value can be converted into corresponding crisp output value. Center of gravity is also one of the factors affecting the Defuzzification. Using Eq. (7) center of gravity denoted below. For all the output crisp value for Fig. 4a rating is 0.904.

$$\text{COG} = \frac{\int_0^1 \mu(t)t dt}{\int_0^1 \mu(t) dt} \quad (7)$$

#### 4.5 Node Architecture in MANET

In a node model a node can have different services to serve. In that based on rating the node can serve the corresponding service. A node can have different services denoted as  $N = (N_{id}, \{S_1, S_2, \dots, S_n\} E_{Node})$  each service can have different attributes such as service ID, service IP address, service output, service input, service response time, service throughput,  $c$  other constraints, Fig. 4 represents node with different services.

### 5 Qos Constraint Service Discovery and Service Composition

#### 5.1 Service Discovery

In the earlier research work they explained about the service discovery [10]. In the MANET's nodes can join and leave on the fly. It is hard to maintain the network structure. Based on dynamic topology they can form the new networks and there is no decentralized system to maintain the nodes in the network. The nodes present in the network are belongs to one of the available category and discovery can repeatedly done based on the nodes can enter into a network else leaves from the network.

#### 5.2 Qos Constraint Service Composition

In this section, the service composition can be done based on the service constraints, considering the service constraints such as energy index value is  $E_{node}$ , and hop count  $h$  distance from one composition initiator to another service provider and as well as services and finally apply the fuzzy logic on the service constraints and normalize

into a single parameter such as rating of a service that can be stored in matrix format as shown below.

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} & \dots & R_{1n} \\ R_{21} & R_{22} & R_{23} & \dots & R_{2n} \\ R_{31} & R_{32} & R_{33} & \dots & R_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ R_{m1} & R_{m2} & R_{m3} & \dots & R_{mn} \end{bmatrix}$$

Each row represents the set of indistinguishable services can performed by distinct nodes and each column represents a set of services performed by a node. First list out the all the services such as  $s_1, s_2, s_3, s_4, \dots, s_n$ , after the execution plan the composition path to be formed. For example a service request  $s_1$  can be formed that to be served by the node  $n_1$  and that the composition plan to be given to the service  $s_2$  present in node  $N_2$ .

Algorithm: service composition algorithm(Rating, n, m)

Procedure ServiceComposition Algorithm(N, n, m)

//Rating[1:n; 1:m] is the rating for a services

//MaxRating[i;j] maximum rated services

//m is the number of nodes in the composition

//n is the number of services

for i <--- 1 to n

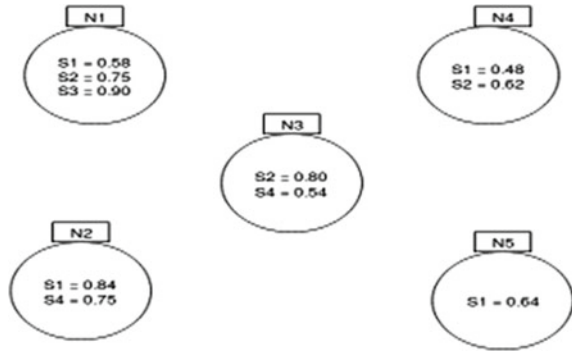
```
{
|   MaxRating[i] ---> 0;
|   for j <--- 1 to m
|   {
|   |   if Rating[j,i] > MaxRating[i] then
|   |   |   MaxRating[i] <--- j
|   |   }
|   }
```

Composition path is MaxRating{1:n};

```
{
|   for i <--- 1 to n-1
|   {
|   |   Execute Service{MaxRating[i];
|   |   Append Input Output;
|   |   create comparison request packet to MaxRating[i+1] Service;
|   |   Hand Over Composition Request( =MaxRating[i+1])
|   |   }
|   }
```

// Transfer composition Results to initiator

**Fig. 5** Nodes with several services and fuzzy rating values



AODV routing protocol  $s_1$  can help to find out the node  $n_2$  when the  $s_2$  service to be present this will continues until the service composition path to be established and the results can be transferred to service compositor. For example let us consider the matrix for representing the rating for different services.

$$\begin{pmatrix} 0.58 & 0.84 & \infty & 0.48 & 0.64 \\ 0.75 & \infty & 0.80 & 0.62 & \infty \\ 0.90 & \infty & \infty & \infty & \infty \\ \infty & 0.75 & 0.54 & \infty & \infty \end{pmatrix}$$

**Node Architecture**

As shown in Fig. 5 the node  $n_1$  consists of three services such as  $s_1, s_2$  and  $s_3$ . Node  $n_2$  consists of two services i.e.,  $s_1$  &  $s_4$ . Node  $n_3$  consists of services  $s_2$  and  $s_4$  &  $n_4$  consists of services  $s_1$  &  $s_2$ . Node  $n_5$  consists of services  $s_1$ . Among these services  $s_1$  can selects node  $n_2$  because that can have maximum rating and  $s_2$  can selects node  $n_3$ , service  $s_3$  can selects node  $n_1$ , and finally service  $s_4$  can selects node  $n_2$ . The composition path to be like  $n_2 \rightarrow n_3 \rightarrow n_1 \rightarrow n_2$  (Fig. 6).

**6 Simulation Results**

Here, we are going to compare and identify the services involved in the composition versus throughput, turnaround time and average packet delay, Table 3 summarizes the simulation set up.



Fig. 6 Composition path

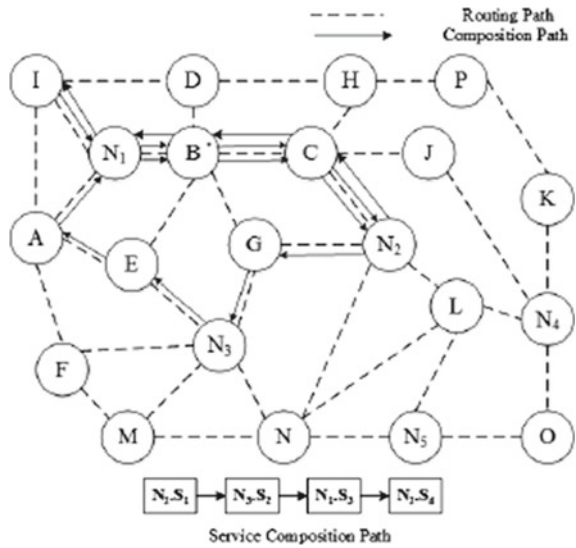


Table 3 Simulation parameters

Parameters	Value
MAC layer	IEEE 802.11
Simulation area (m <sup>2</sup> )	1000 m * 1000 m
Simulation time	60 s
Number of nodes	25
Bandwidth	2 Mbps
Node mobility speed	0–60 m/s
Traffic flow	CBR
Packet size	512 bytes
Transmission range	250 m

### 6.1 Comparison Among Average Packet Delay and Number of Services

Figure 7 shows the comparison among the number of required services and average packet delay. Here we can know that when the stable nodes are present the average packet delay decreases. The proposed method increases the network life time and decreases the average packet delay by involving the energy constraint services and least hop-count when providing the service.

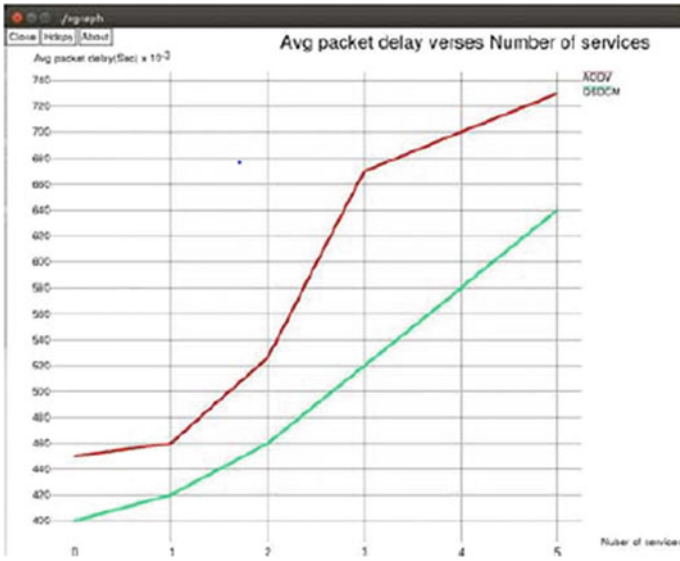


Fig. 7 Average packet delay versus number of services

### 6.2 Comparison Among Throughput and Number of Services

In previous work throughput will be decreases when the number of services increased. In the proposed work throughput will be increased as compared to AODV protocol. In proposed work node can serve more than one services that results overhead will be decreased and throughput will increased (Fig. 8).

### 6.3 Comparison Among Energy and Number of Services

In Fig. 9 shows comparison among energy and number of services among the proposed work and AODV protocol. In proposed method it consumes less energy because it uses service constraints and results increases network life time of a network.

### 6.4 Turn-Around Time Vs Number of Services

In Fig. 10, shows comparison among turn-around time and number of required services presented in the composition. As compared to AODV, proposed work can provide the better results and selects the maximal optimal service by using fuzzy approach.

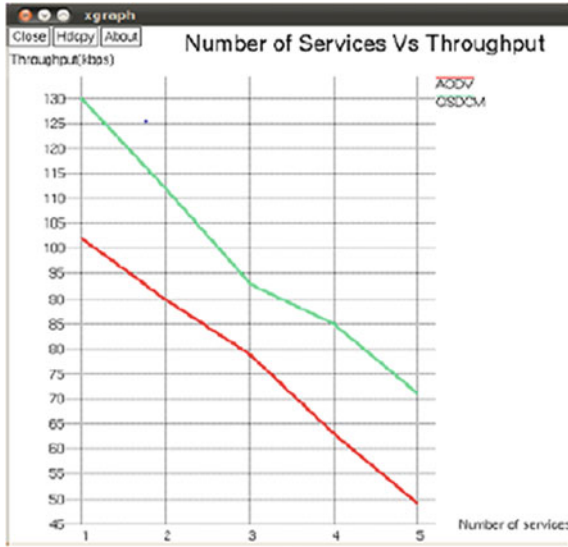


Fig. 8 Comparison among throughput and number of services

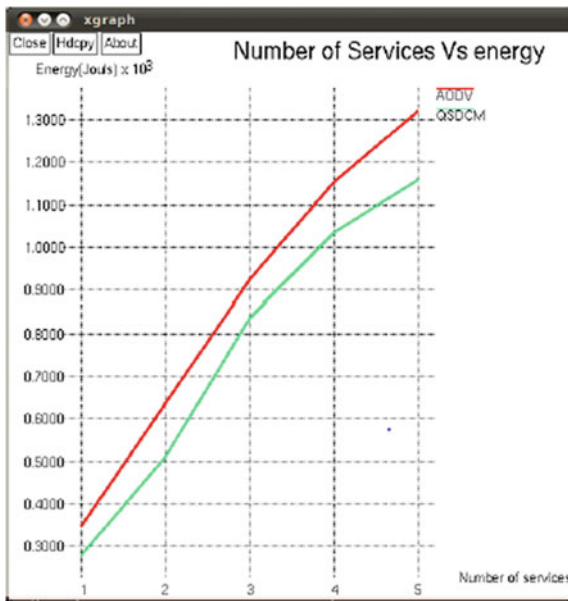


Fig. 9 Comparison among energy and number of services

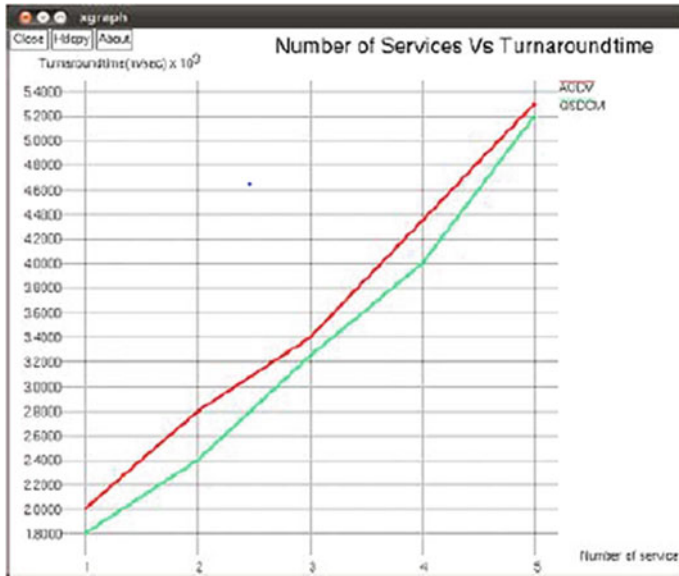


Fig. 10 Turnaround time versus number of services

## 7 Conclusion and Future Work

In existing service composition in the MANET's they won't consider QoS constraints at the node and service layer. In proposed system we are considering response time, throughput as a service QoS constraints and hop count and energy consumption as the node QoS constraints. By considering these service constraints we apply the fuzzy rule on these service constraints able to evaluate rating of a service there by selects maximal optimal service during the service composition for the dynamic networks. Our approach is highly adaptable and scalable in providing the service. Simulation results represents that our proposed approach is better performance than AODV protocol. In future work we are going to consider fault tolerance to increase reliability of a service and to minimize the service failure rates.

## References

1. Y. Wang, R. Chen, J.H. Cho, A. Swami, K.S. Chan, Trust-based service composition and binding with multiple objective optimization in service-oriented mobile ad hoc networks. *IEEE Trans. Ser. Comput.* **10**(4), 660–672 (2017)
2. I.-R. Chen, J. Guo, D.-C. Wang, J.J.P. Tsai, H. Al-Hamadi, I. You, Trust-based service management for mobile cloud IoT systems. *IEEE Trans. Netw. Serv. Man.* **16**(1), 246–263
3. W. Muhamad, Suhardi, Y. Bandung, A research challenge on mobile and cloud service composition. in *Information Technology Systems and Innovation (ICITSI) 2018 International*

- Conference on, pp. 347–352 (2018)
4. D. Chakraborty, Y. Yesha, A. Joshi, A distributed service composition protocol for pervasive environments. in *IEEE Wireless Communications and Networking Conference*, pp. 2575–2580 (2004)
  5. I.-R. Chen, J. Guo, D.-C. Wang, J.J.P. Tsai, H. Al-Hamadi, I. You, Trust as a service for IoT service management in smart cities. in *High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) 2018 IEEE 20th International Conference on*, pp. 1358–1365 (2018)
  6. R.A. Rehman, B.S. Kim, LOMCF: forwarding and caching in named data networking based MANETs. *IEEE Trans. Veh. Netw*
  7. C. Wu, M. Gerla, M. van der Schaar, Social norm incentive for network coding in MANETs. *IEEE/ACM Trans. Netw*
  8. M.S. Khan, D. Midi, M.I. Khan, E. Bertino, Fine-grained analysis of packet loss in MANETS. *IEEE j. Mag.* **5**, 7798–7807 (2017)
  9. A. Taha, R. Alsaqour, M. Uddin, M. Abdelhaq, T. Saba, Energy efficient multipath routing protocol for mobile ad-hoc network using the fitness function. *IEEE Access* **5**, 10369–10381 (2017)
  10. F. Bao, R. Chen, M. Chang, J.H. Cho, Hierarchical trust management for wireless sensor networks and its application to trust-based routing and intrusion detection. *IEEE Trans. Netw. Ser. Manage.* **9**(2), 169–183 (2012)

# Inferential Procedures for Testing Assumptions on Observations for Applications of Biometric Techniques



M. Naresh, B. Sarojamma, P. Srivyshnavi, G. Madhusudan, P. Vishnupriya, and P. Balasiddamuni

**Abstract** Biometrical Techniques are often used in Genetic Statistics involving plants and animal studies for assessing their genetic potential in selection trails for genetic material improvement. For such purposes, genetic parameters namely means, variances, variance components, heritability parameters, genotypic correlations etc., have been often estimated by using genetic statistical methods. In the Biometrical research analysis, the applications of most of the advanced experimental statistical tools based on certain crucial assumptions such as the assumptions of independence, homoscedasticity of observations on study variable and assumption of normality of observations in the data. Departures from these assumptions may lead to biased and inconsistent estimators; and incorrect conclusions. Thus, the Biostatistician has to test these assumptions on observations rather than to presume that they are correct. In the present article, an attempt has been made by developing test procedures for testing hypotheses about population's symmetry and population's kurtosis by using some modified Beta measures. Further, a test for normality of errors in linear regression model has been developed by using modified Fisher's  $g$ -statistics.

**Keywords** MMSE estimators · Symmetry and kurtosis measures · Normality of errors

## 1 Introduction

In much of the theoretical and applied research on applications of Biometric techniques, most of the advanced statistical inferential methods are based on certain crucial assumptions of independence, homoscedasticity and normality of observations in the data. Violations of these assumptions may lead to several problems in the applications of Biometric Techniques. Consequently, one may obtain biased and

---

M. Naresh (✉) · B. Sarojamma · G. Madhusudan · P. Vishnupriya · P. Balasiddamuni  
Department of Statistics, S.V. University, Tirupati, India  
e-mail: [naresh.ygr@gmail.com](mailto:naresh.ygr@gmail.com)

P. Srivyshnavi  
Department of CSC, S.P.M.V.V. Engineering College, Tirupati, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_33](https://doi.org/10.1007/978-3-030-46939-9_33)

inconsistent estimators for parameters resulting incorrect conclusions from applications of inferential procedures. Thus, there is necessity to test these assumptions rather than to presume that they are correct. In the applications of Biometric Techniques based on Applied Regression Analysis, the presence of non normal errors in the linear regression models disturbs optimum properties of best estimators such as Ordinary Least Squares (OLS) estimators of the parameters. In the case of Analysis of Variance (ANOVA) technique, one has to test the normality assumption of observations before its application. In practice, it is desired to test the hypothesis that a sample came from a population, whose distribution is normal distribution. The Beta measures of skewness and kurtosis of the population,  $\sqrt{\beta_1}$  and  $\beta_2$  respectively can be used as measures for evaluating Normality of population. The corresponding Sample Statistics measures  $\sqrt{b_1}$  and  $b_2$  can be considered as descriptive and inferential measures for evaluating Normality of the population.

## 2 Moments Coefficients as Measures of Skewness and Kurtosis

Write the  $p$ th central moment about the population arithmetic mean ( $M$ ) for a population consists of  $N$  units as  $M_p = \frac{\sum_{i=1}^N (X_i - \mu)^p}{N}$ ,  $p = 1, 2, 3, 4, \dots$ , where  $\mu =$  Population Mean,  $N =$  Size of Population. For any population distribution,  $M_1 = \frac{\sum (X - \mu)}{N} = 0$  and  $M_2 = \frac{\sum (X_i - \mu)^2}{N}$  is the population variance  $\sigma^2$ . The various coefficients based on moments to measure Skewness and Kurtosis are given by:

(a)  $\alpha$ -Coefficients:

$$\alpha_1 = \frac{M_1}{\sigma} = \frac{M_1}{\sqrt{M_2}} = 0, \alpha_2 = \frac{M_2}{\sigma^2} = \frac{M_2}{M_2} = 1,$$

$$\alpha_3 = \frac{M_3}{\sigma^3} = \frac{M_3}{M_2^{3/2}}, \alpha_4 = \frac{M_4}{\sigma^4} = \frac{M_4}{M_2^2}$$

Here  $\sigma$  is population Standard Deviation.

(b)  $\beta$ -Coefficients:

$$\beta_1 = \frac{M_3^2}{M_2^3} = \alpha_3^2, \sqrt{\beta_1} = \frac{M_3}{M_2^{3/2}} = \alpha_3, \beta_2 = \frac{M_4}{M_2^2} = \alpha_4$$

(c)  $\gamma$ -Coefficients:

$$\gamma_1 = \sqrt{\beta_1} = \alpha_3, \gamma_2 = \beta_2 - 3 = \frac{M_4 - 3M_2^2}{M_2^2}.$$

**Remark** For a Symmetrical distribution, all the odd order moments about Mean (Central Moments) vanish. i.e.,  $M_1, M_3, M_5, \dots$ , are all equal to zero.

### 3 Different Measures of Skewness

Different measures of Skewness are given by:

- (a)  $\alpha_3$  **Coefficient of Skewness:**  $\alpha_3 = \frac{\mu_3}{\sigma^3} = \sqrt{\beta_1} = \gamma_1$   
Here,  $\alpha_3$  may be either positive or negative according to algebraic sign of  $\mu_3$ .
- (b) **Coefficient of Skewness Based on  $\beta_1$  and  $\beta_2$ :**  
 $S_M = \frac{\sqrt{\beta_1(\beta_2+3)}}{2(5\beta_2-6\beta_1-9)}$ , Limits of  $S_M$ :  $-3 \leq S_M \leq 3$

**Remark** For a Skewed distribution, Mode can be computed by using Karl Pearson's Coefficient of Skewness. i.e., Mode = Arithmetic Mean  $- \sigma S_M$

- (c) **Karl Pearson's Coefficient of Skewness:**  
 $S_k = \frac{\text{Arithmetic Mean}-\text{Mode}}{\text{Standard Deviation}}$  or  $S_k = \frac{3[\text{Arithmetic Mean}-\text{Median}]}{\sigma}$ ,  
Limits of  $S_k$ :  $-3 \leq S_k \leq 3$
- (d) **Bowley's Quartiles Coefficient of Skewness:**  
 $S_Q = \frac{(Q_3-Q_2)-(Q_2-Q_1)}{Q_3-Q_1}$  or  $S_Q = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$ , where  $Q_1, Q_2$  and  $Q_3$  are first, second and third Quartiles of population. Limits of  $S_Q$ :  $-1 \leq S_Q \leq 1$
- (e) **Kelly's Percentiles Coefficients of Skewness:**
  - (i)  $S_P = \frac{(P_{90}-P_{30})-(P_{50}-P_{10})}{2}$  or  $S_P = \frac{P_{90}+P_{10}-2P_{50}}{2}$
  - (ii) Modified  $S_P = \frac{P_{90}+P_{10}-2P_{50}}{P_{90}-P_{10}}$ , where,  $P_i, i = 10, 50, 90$ , is the  $i$ th percentile of population. Limits of  $S_P$ :  $-1 \leq S_P \leq 1$
- (f)  $\gamma_1$  **Coefficient of Skewness:**  
 $\gamma_1 = \sqrt{\beta_1}$ .

### 4 Different Measures of Kurtosis

The different measures of kurtosis are given by:

- (a) **Karl Pearson's Moments Coefficient of Kurtosis:**  
 $\beta_2 = \frac{M_4}{\sigma^4} = \frac{M_4}{M_2^2} = \alpha_4$ ,  $\beta_2 > 3 \Rightarrow$  Leptokurtosis,  $\beta_2 = 3 \Rightarrow$  Mesokurtosis  
 $\beta_2 < 3 \Rightarrow$  Platykurtosis
- (b) **Moors [1] Octiles Coefficient of Kurtosis:**  
 $R_0 = \frac{(O_7-O_5)+(O_3-O_1)}{O_6-O_2} \Rightarrow R_0 = \frac{(O_7-O_5)+(O_3-O_1)}{Q_3-Q_1}$ , where  $O_1 =$  First octile = 12.5th percentile =  $P_{12.5}$ ;  $O_3 =$  Third octile = 37.5th percentile =  $P_{37.5}$ ;  $O_5 =$  Fifth octile = 62.5th percentile =  $P_{62.5}$ ;  $O_7 =$  Seventh octile = 87.5th percentile =  $P_{87.5}$ ;  $O_2 = Q_1 = P_{25}$ ,  $O_4 = Q_2 = P_{50}$ ;  $O_6 = Q_3 = P_{75}$ ; Here,  $R_0 = 0 \Rightarrow$  Extreme platy Kurtosis,  $R_0 = 1.233 \Rightarrow$  Mesokurtosis  $R_0 = \infty \Rightarrow$  Extreme Leptokurtosis.
- (c) **Kelly's Percentiles Measure of Kurtosis:**  
 $R_p = \frac{P_{75}-P_{25}}{P_{90}-P_{10}}$ ,  $R_p > 0.26315 \Rightarrow$  Platy Kurtosis,  $R_p < 0.26315 \Rightarrow$  Leptokurtosis,  $R_p = 0.26315 \Rightarrow$  Mesokurtosis.



(d)  $\gamma_2$  **Coefficient of Kurtosis:**

$\gamma_2 = \beta_2 - 3$ ,  $\gamma_2 > 0 \Rightarrow$  Leptokurtosis,  $\gamma_2 = 0 \Rightarrow$  Mesokurtosis,  $\gamma_2 < 0 \Rightarrow$  Platy Kurtosis.

### 5 Fisher [2], Pearson [3] Measures of Symmetry and Kurtosis

Suppose that  $\sqrt{\beta_1}$  and  $\beta_2$  as population parameter measures of Symmetry and Kurtosis respectively. The sample estimates of  $\sqrt{\beta_1}$  and  $\beta_2$  based on Fisher’s g-statistics are given by

- (i)  $\sqrt{b_1} = \sqrt{\hat{\beta}_1} = \frac{(n-2)g_1}{\sqrt{n(n-1)}}$ . Here,  $\sqrt{b_1} = 0 \Rightarrow$  Symmetrical distribution
- (ii)  $b_2 = \hat{\beta}_2 = \frac{(n-2)(n-3)g_2}{(n+1)(n-1)} + \frac{3(n-1)}{n+1}$ . Here,  $b_2 = \frac{3(n-1)}{n+1} \Rightarrow$  Mesokurtosis.

### 6 Minimum Mean Square Error (MMSE) Estimator for Variance Parameter of Normal Population

Suppose that  $(X_1, X_2, \dots, X_n)$  be a random sample drawn from Normal Population  $N(\mu, \sigma^2)$ . Consider the sample variance  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$  based on  $n$  random observations  $x_1, x_2, \dots, x_n$  from  $N(\mu, \sigma^2)$ . It is known that,  $s^2$  is an unbiased estimator of  $\sigma^2$  i.e.,  $E[s^2] = \sigma^2$ . Also,  $\left[\frac{(n-1)s^2}{\sigma^2}\right] \sim \chi^2_{(n-1)}$ . One may obtain

$$\text{Var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = \text{Var}(\chi^2_{(n-1)}) = 2(n-1) \Rightarrow \text{Var}(s^2) = \frac{2\sigma^4}{n-1}$$

Now, consider,

$$\begin{aligned} E[Cs^2 - \sigma^2]^2 &= E[Cs^2 - C\sigma^2 + C\sigma^2 - \sigma^2]^2 = E[C(s^2 - \sigma^2) - \sigma^2(1 - C)]^2 \\ &= C^2\text{Var}(s^2) + \sigma^4(1 - C)^2 \Rightarrow E[Cs^2 - \sigma^2]^2 = \left[\frac{2C^2\sigma^4}{n-1} + (1 - C)^2\right]\sigma^4 \end{aligned}$$

Then attains the minimum when  $C = \frac{(n-1)}{n+1}$ , the minimum value being  $\left[\frac{2\sigma^4}{n+1} < \frac{2\sigma^4}{n-1}\right]$ . One may obtain, for all  $n$  and  $\sigma^2$ ,  $E\left[\frac{\sum(x_i - \bar{x})^2}{n+1} - \sigma^2\right]^2 <$

$E\left[\frac{\sum(x_i - \bar{x})^2}{n-1} - \sigma^2\right]^2$  and  $s_1^2 = \frac{\sum(x_i - \bar{x})^2}{n+1}$  is a biased estimator for  $\sigma^2$ . Under minimum Mean Square error criterion,  $s_1^2 = \frac{\sum(x_i - \bar{x})^2}{n+1}$  is biased estimator for  $\sigma^2$  and is better estimator than an unbiased estimator  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ .

### 7 Modified Estimates of Population Moments; Modified g-Statistics and Modified Beta Measures

Consider the Minimum Mean Square Estimate of Variance parameter based on random sample of  $n$  observations as  $s_1^2 = \frac{\sum(x - \bar{x})^2}{n+1}$ . The modified estimators of central moments  $\mu_2, \mu_3$  and  $\mu_4$  are given by:

$$\widehat{M}'_2 = \frac{\sum(x - \bar{x})^2}{n + 1} = s_1^2 = \text{MMSE estimator}$$

$$\widehat{M}'_3 = \frac{\sum(x - \bar{x})^3}{n + 1} = \frac{n \sum x^3 - 3 \sum x \sum x^2 + 2(\sum x)^3 / n}{n(n + 1)}$$

$$\widehat{M}'_4 = \frac{(n + 1)n(n - 1) \sum(x - \bar{x})^4 - 3[\sum(x - \bar{x})^2]^2}{n(n - 1)}$$

or  $\widehat{M}'_4 = \frac{(n^3 + n^2) \sum x^4 - 4(n^2 + n) \sum x^3 \sum x - 3(n^2 - n) [\sum x^2]^2 + 12n [\sum x^2] [\sum x]^2 - 6[\sum x]^4}{(n+2)(n+1)n(n-1)}$

or  $\widehat{M}'_4 = \left[ \frac{(n+1) \left( n \sum x^4 - 4 \sum x \sum x^3 + 6 [\sum x]^2 \left( \frac{\sum x^2}{n} \right) - \frac{3 [\sum x]^4}{n^2} \right)}{(n+1)n(n-1)} \right] - \left[ \frac{3 \left( \sum x^2 - \frac{(\sum x)^2}{n} \right)}{n(n-1)} \right]$

The modified g-statistics estimators are given by

$$g'_1 = \frac{\widehat{M}'_3}{\sqrt{(s_1^2)^3}} \text{ and } g'_2 = \frac{\widehat{M}'_4}{\sqrt{(s_1^2)^2}}$$

The modified sample estimates of Beta measures of Skewness and Kurtosis based on g-statistics are given by

- (i)  $\sqrt{b'_1} = \sqrt{\widehat{\beta}'_1} = \frac{(n-2)g'_2}{\sqrt{n(n-1)}}$  Here,  $\sqrt{b'_1} = 0 \Rightarrow$  Symmetrical distribution
- (ii)  $b'_1 = \widehat{\beta}'_2 = \frac{(n-2)(n-3)g'_2}{(n+1)(n-1)} + \frac{3(n-1)}{n+1}$  Here,  $b'_2 = \frac{3(n-1)}{n+1} \Rightarrow$  Mesokurtosis.

## 8 Testing Hypothesis About Population’s Symmetry and Population’s Kurtosis

According to the tests for the normality suggested by  $D'$  Agostino et al. [4], the modified tests for the symmetry and kurtosis of the population are developed by using the proposed modified Beta measures of Symmetry and Kurtosis.

### 8.1 Test for the Population’s Symmetry

One may state the two tailed hypothesis about the symmetry of the population distribution as  $H_0: \sqrt{\beta_1} = 0$  i.e. The population is specified by Symmetrical distribution against  $H_1: \sqrt{\beta_1} \neq 0$  i.e. The population is not specified by Symmetrical distribution. To test the null hypothesis, one may use the following modified test procedure:

**Step (1):** First, compute  $\sqrt{b'_1}$  by using  $g'_1$  as  $\sqrt{b'_1} = \frac{(n-2)g'_1}{\sqrt{n(n-1)}}$  where  $g'_1 = \frac{\widehat{M}_3'}{\sqrt{(s_1^2)^3}}$

**Step (2):** Calculate the modified Beta function  $f(\sqrt{b'_1})$  as

$$f(\sqrt{b'_1}) = \frac{\sqrt{b'_1} \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}}{\sqrt{2}l(W_2 - 1)}, \text{ where, } W_2 = \sqrt{2(W_1 - 1)} - 1,$$

$$W_1 = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

**Step (3):** Compute the modified test statistic for testing the null hypothesis of population’s Symmetry as

$$Z_{g'_1} = W_3 \text{Ln} \left[ f(\sqrt{b'_1}) + \sqrt{[f(\sqrt{b'_1})]^2 + 1} \right] \text{ where } W_3 = \frac{1}{\sqrt{\text{Ln}(W_2)} / 2}$$

**Step (4):** Compare the calculated value of  $Z_{g'_1}$  with its critical value and draw the inference accordingly. Here,  $Z_{g'_1} \sim N(0, 1)$ .

### 8.2 Test for the Population’s Kurtosis

One may state the null hypothesis about the Mesokurtosis of the population distribution as  $H_0: \beta_2 = 0$  i.e. The population is specified by Mesokurtic distribution against  $H_1: \beta_2 \neq 0$  i.e. The population is not specified by Mesokurtic distribution. The following test procedure based on the modified measures of Kurtosis may be used to test the null hypothesis:

**Step (1):** Compute the modified g-statistics function as

$$f(g'_2) = \frac{(n-2)(n-3)|g'_2|}{(n+1)(n-1)\sqrt{T_1}} \text{ where } T_1 = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

**Step (2):** Calculate a modified test function  $\phi$  as

$$\phi = \frac{1 - \frac{2}{T_2}}{1 + f(g'_2)\sqrt{\frac{2}{T_2-4}}} \text{ where, } T_2 = 6 + \frac{8}{T_3} \left[ \frac{2}{T_3} + \sqrt{1 + \frac{4}{T_3^2}} \right]$$

$$T_3 = \frac{6(n^2 - 5n + 2)}{(n + 7)(n + 9)} \sqrt{\frac{6(n + 3)(n + 5)}{n(n - 2)(n - 3)}}$$

**Step (3):** The modified test statistic for testing the null hypothesis of Mesokurtic population distribution as

$$Z_{g'_2} = \left[ \frac{1 - \frac{2}{9T_2} - \frac{3\phi}{\sqrt{2/9T_2}}}{\sqrt{2/9T_2}} \right] \sim N(0, 1), \text{ Compare the calculated value of } |Z_{g'_2}| \text{ with it's critical value and draw the inference accordingly.}$$

### 9 A Modified Test for the Normality of Population

To test for the normality of the population, the modified test statistic based on the developed modified test statistics is given by  $U^2(g'_1, g'_2) = (Z^2_{g'_1} + Z^2_{g'_2})$ , where  $Z^2_{g'_1}$  and  $Z^2_{g'_2}$  are given in Sects. 8.1 and 8.2 respectively. Here,  $U^2(g'_1, g'_2) \sim \chi^2_2$ . Compare the calculated value of test statistic  $U^2(g'_1, g'_2)$  with it's critical value and draw the inference accordingly.

**Remark** If the hypothesis of normality of population is rejected, then the non normality due to either departure from Symmetry or departure from Mesokurtosis may be tested by using the modified test procedures separately or both given in Sects. 8.1 and 8.2.

### 10 A Modified Wald Test for Normality of Errors in Linear Regression Model

Consider the standard linear statistical model  $y_i = X'_i\beta + \varepsilon_i, \quad i = 1, 2, \dots, n$  where  $X_1, X_2, \dots, X_n$  are  $(k \times 1)$  vectors;  $y_i$  is  $i$ th observation on dependent variable; which  $\beta$  is  $(k \times 1)$  vector;  $\varepsilon_i$ 's are random disturbances which follow  $N(0, \sigma^2)$ . Suppose that  $\hat{\beta}$  be the Ordinary Least Squares (OLS) estimator of parametric vector  $\beta$ . The OLS residuals  $e'_i$ 's are given by

$e_i = y_i - \hat{y}_i = X'_i\beta + \varepsilon_i - (X'_i\hat{\beta})$  or  $e_i = \varepsilon_i - X'_i(\hat{\beta} - \beta)$  where the OLS estimator is given by  $\hat{\beta} = (X'X)^{-1}X'Y$ . Here,  $Y$  is  $(n \times 1)$  vector,  $X$  is  $(n \times k)$  matrix and  $\hat{\beta}$  is  $(k \times 1)$  vector of OLS estimators. It is known that the sample of OLS residuals converges to the sample of true errors. The modified OLS residual estimators of 2nd, 3rd and 4th Central moments of residuals are defined as

$$\widehat{M}'_{2R} = \frac{\sum_{i=1}^n e_i^2}{n+1}, \widehat{M}'_{3R} = \frac{\sum e_i^3}{n+1}, \widehat{M}'_{4R} = \frac{(n+1)n(n-1) \sum e_i^2 - 3(\sum e_i^2)^2}{n(n-1)}$$

By defining the modified sample estimators of Beta measures of Skewness and Kurtosis as  $\sqrt{b'_{1R}} = \frac{(n-2)g'_{1R}}{\sqrt{n(n-1)}}$  and  $b'_{2R} = \frac{(n-2)(n-3)}{(n+1)(n-1)}g'_{2R} + \frac{3(n-1)}{n+1}$  the Modified Wald test statistic for testing the normality of residuals is given by  $Q'_R = n \left[ \frac{b'_{1R}}{b} + \frac{(b'_{2R}-3)^2}{2n} \right]^{Asy} \sim \chi^2_2$  where,  $g'_{1R} = \frac{\widehat{M}'_{3R}}{\sqrt{(\widehat{M}'_{2R})^3}}$ ,  $g'_{2R} = \frac{\widehat{M}'_{4R}}{(\widehat{M}'_{2R})^2}$ .

### 11 Conclusions

Most of the Biometrical techniques can be validly applied under certain crucial assumptions of independence, homogeneity of variances and normality of observations. Because of the possible consequences of departure of normality of observations, one has to test for the normality of there data. In the present study, some modified tests for symmetry and the normality of population have been developed by using Minimum Mean Square Error estimators of population moments. A modified Wald test for normality of errors in linear regression model also has been developed. The various tests proposed for testing normality of errors in linear regression model can be further studied by replacing OLS residuals by other types of residuals such as Best Linear Unbiased Scalar (BLUS), Recursive, Internally/Externally/Studentized, Predicted, Abrahamse and Koerts, Best Augmented Unbiased with Scalar Matrix (BAUS), Independent step-wise, Uncorrelated, Restricted Least Squares (RLS), Standardized, Weighted and Generalized Least Squares (GLS) residuals.

### References

1. E.S. Pearson, R.B. D'Agostino, K.O. Bowman, Tests for departure from normality: comparison of powers. *Biometrika* **64**, 231–246 (1977)
2. D.S. Moore, Tests of chi-squared type, in *Goodness-of-Fit Techniques*, eds. by R.B. D'Agostino, M.A. Stephens (Marcel Dekkar, New York, 1986), pp. 63–95
3. H.C. Thode Jr, *Testing for Normality* (Marcel Dekkar, New York, 2002), pp. 479
4. R.A. Fisher, *Statistical Methods for Research Workers*, 1st edn. (Oliver and Boyd, Edinbergh, Scotland, 1925)
5. K.D. Bowman, L.R. Shenton, Omnibus contours for departures from normality based on  $\sqrt{b_1}$  and  $b_2$ . *Biometrika* **62**, 243–250 (1975)
6. R.B. D'Agostino, A. Belanger, R.B. D'Agostino Jr, A suggestion for using powerful and informative tests of normality. *J. Am. Stat. Assoc.* **44**, 316–321 (1990)
7. R.B. D'Agostino, E.S. Pearson, tests for departure from normality; empirical results for the distributions of  $b_2$  and  $\sqrt{b_1}$ . *Biometrika*, **60**, 613–622 (1973)
8. R. Groeneveld, G. Meeden, Measuring skewness and kurtosis. *Statistician* **33**, 391–399 (1984)
9. J.J.A. Moors, A quantile alternative for kurtosis. *Statistician* **37**, 25–32 (1988)

10. M. Naresh, Advanced statistical techniques for agricultural research. Unpublished Ph.D. thesis in Statistics, S.V. University, Tirupati, Andhra Pradesh, 2017
11. K. Pearson, Contributions to the mathematical theory of evolution. Philos. Trans. Roy. Soc. Lond. **91**, 343–358 (1895)
12. M.M. Rahman, Z. Govindarajulu, A modification of the test of shapiro and wilk for normality. J. Appl. Statist. **24**, 219–236 (1997)
13. K. Vijaya Kumar, P. Balasiddamuni et al., *Testing Normality in Linear Statistical Models* (Lap Lambert Academic, Germany, 2013). ISBN: 978-3-659-50283-5
14. H.J. Zar, *Biostatistical Analysis*, 5th edn. (Prentice Hall, Upper Saddle River, NJ, 2009)

# Smart Crop Suggester



N. Usha Rani and G. Gowthami

**Abstract** Technology plays key role in every sector. Technical support is important for making decisions in various fields such as medicine, education, commerce, marketing, agriculture etc. The Smart Crop Suggester is android based application which will assist farmer to choose better crop to yield high production. The Smart Crop Suggester application provides a user friendly interface to farmers for getting better crop suggestion suitable for place, season, soil type and rainfall range from the analysis of past year agriculture data. Smart Crop Suggester application helps farmer to make better choice. Smart Crop Suggester also focuses on financial status of the farmer to give crop recommendations.

**Keywords** Android application · Crop suggestion

## 1 Introduction

The main occupation in our country is agriculture. In Andhra Pradesh about 62% of the people are on agriculture. Agriculture is the major source of income in India as well as in Andhra Pradesh. Important crops grown in Andhra Pradesh are rice, Jowar, Bajra, Maize, Cotton, Sugarcane, Pulses etc.

---

N. Usha Rani (✉) · G. Gowthami

Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Tirupati, AP, India  
e-mail: [usha552@yahoo.com](mailto:usha552@yahoo.com); [usharani.ur@gmail.com](mailto:usharani.ur@gmail.com)

G. Gowthami

e-mail: [gavalagowthami1995@gmail.com](mailto:gavalagowthami1995@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_34](https://doi.org/10.1007/978-3-030-46939-9_34)

A mission “Sustainable Agriculture Production with minimum cost of cultivation, eventually enhancing the return on income to the farmer” is being implemented by the Government. As a part of the mission Government is undertaking several schemes for the farmer such as free power supply for 9 h, subsidy for seeds, crop loans in subsidized interest rate etc.,. Andhra Pradesh bags the first position in providing agricultural loans from commercial and cooperative banks. At present there are 13 districts in the state of Andhra Pradesh, there are several types of crops grown in different districts of Andhra Pradesh and soil types vary from district to district. All types of soils are not suitable for all types of crops. It is difficult for the farmer to identify which type of soil is suitable for which crop. Agriculture in Andhra Pradesh is facing a lot of problems due to ever increasing population, it is also difficult to meet the food needs of the people. Farmers need to concentrate on increasing food production by choosing suitable season and suitable soil for crops. This application provides a platform for the farmers to make better choice. As the technology is growing very rapidly and everybody is accessing smartphones this application has been developed as an android application.

“Use of technology today and tomorrow” focused in calculating the amount of grass in the field using mobile application. It saves time and money for the farmer. Farmers know amount of grass that is available for feeding their animals. Technology is helping farmers to face real time problems in a efficient way. Technology is helping consumer to place order online and farmer will deliver fresh products to the consumer in time. This will save the farmer money and cut out mediators who will sell products from the farmer with high cost. Every farmer is using technology for their requirement. This is useful for the farmers to choose fertilizers, production and marketing [1].

Agriculture has benefited farmer from the information exchange. Information refers to seasonal change, demand, cultivation. Weather related information is very helpful for the farmers to defend from the losses that may occur due to in appropriate weather conditions [2].

In monitoring and controlling crop irrigation systems mobile technology is playing crucial role. Using mobile it is easy to control irrigation system instead of going to field. This is possible because of moisture sensors present in the soil. Crop sensors are used in farms now a days in helping farmer apply fertilizers in a effective manner. They help in reducing the leaching potential and groundwater run off. Sensors will tell application equipment how much fertilizers crop may need in reality. Based on the amount of light reflected back to sensor, Optical sensors are used to see amount of fertilizer a plant may need [3].

## 2 Literature Survey

Kisanseva is a mobile application which assists farmer by providing relevant information to sell their crops. This application provides direct link between farmer and the customer. Crops in geographical area can be searched. It helps the farmer in



giving better value to the products. People can search crops grown in that particular area and purchase them according to the committee prices. It also gives information about product quality, quantity and price [4].

E agro application is a platform which assists farmer in taking decision regarding selection of pesticides, fertilizers and time to do particular farming actions. For each type of crop the fertilizer schedule will be registered in this application. Also crop sowing date will be mentioned, based on that date farmer will get notifications regarding application of fertilizers. Suggestion will be given to the farmer based on soil type and geographical location [5].

Krishi Ville application assists farmers in farming activities. This application provides agricultural news updates, agricultural commodities, weather forecast updates. This application takes into consideration Indian farming [6].

Smart cultivation partner is an application which helps Sri Lanka farmers who do not have technical and technological skills. People lack awareness regarding markets and actual prices to their products. This application provides good profits by direct communication between, farmer-to-farmer and farmer-to-merchant. This system fill gap between farmers and merchants [7].

E agriculture information management system is an android application which provide awareness, usage and perception in e-agriculture. Data is collected from farmers using statistical techniques to provide awareness. This application acts as a platform for the farmers to support agricultural products marketing [8].

### 3 Smart Crop Suggester

Right crop at right season is important for better yielding. The present android application will be helpful for the farmers to choose the crop. The present Smart Crop Suggester (SCS) will suggest appropriate crops based on the season, type of soil and water sources. In this work, it also focused on the financial status of the farmer which he/she can afford for that crop. If the water resources are not there, it will take the range of rainfall from the past years data and it will suggest the better crop. In this way, SCS helps by giving better suggestion for farmer to get better crop.

#### Design

See Fig. 1.

#### System requirements

Hardware requirements:

- Microsoft Windows 7/windows 8/windows 10 (32-bit or 64-bit)
- Minimum 3 GB RAM, 8 GB RAM is recommended
- Minimum 2 GB of disk space, 4 GB is recommended
- Intel core i3 minimum, i5/i7.

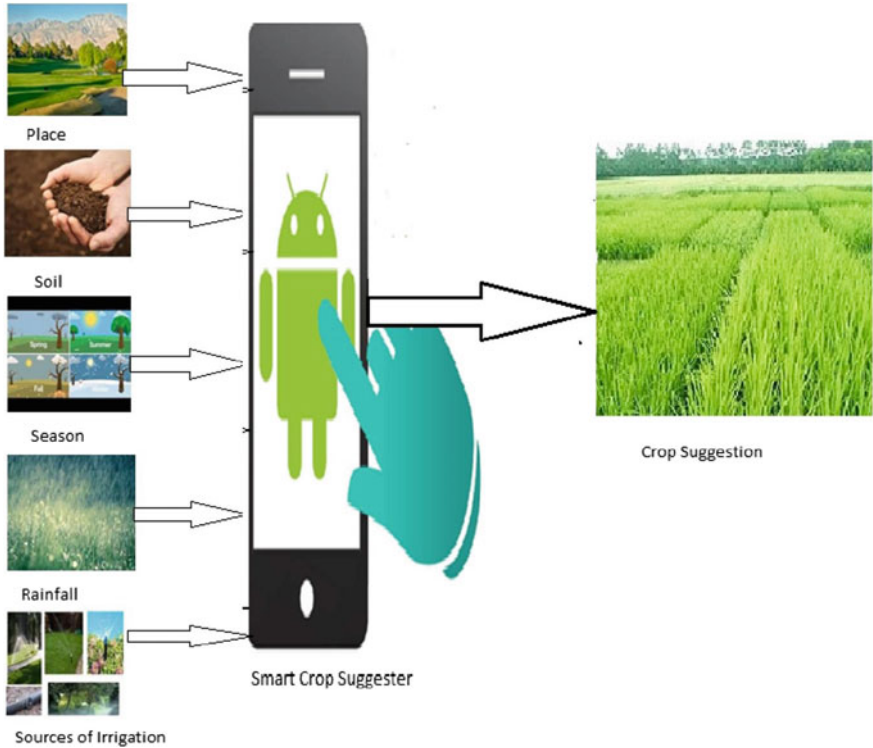


Fig. 1 Design of smart crop suggester application

Software requirements:

- Java sdk
- Android IDE with sdk bundle
- Genymotion
- Oracle VM Virtual Box.

**Limitations to run this application**

- This application won't work properly for Api levels below 26
- Since this is a third party application, it won't work properly in some Smartphones like Redmi, realme etc.
- This application works better from the android versions 8.0 (Oreo) and above.

**Steps to run the application**

Create an Android project:

To create new Android project, follow these steps:

- Latest version of **Android Studio** need to be installed.
- **Welcome to Android Studio** window opens, Then we need to click **Start a new Android Studio project**.
- Go to **Choose your project** window, select **Empty Activity** and click on **Next**.
- Go to **Configure your project window**, then complete the following steps.
- Give “app name” in the Name field.
- Give “com.example.appname” in the **Package name** field.
- From the Language drop-down menu, select **Java**.
- Leave the other options as they are.
- Click on **Finish**.
- Write code.

### Run android application

To Run on a real device

Set up device in the following steps:

1. With USB cable connect phone to the development machine. Appropriate USB drive need to be installed.
2. Perform the following steps to enable USB debugging in the Developer options window:
  - (a) Go to Settings app.
  - (b) Select **About phone** by scrolling to the bottom of the device.
  - (c) Tap **Build number** seven times by scrolling to the bottom.
  - (d) Scroll to the bottom, and tap **Developer options**.
  - (e) In the Developer options window, enable **USB debugging**.

### Run the app on an emulator as follows

1. Create an Android Virtual Device (AVD) in Android Studio.
2. In the toolbar, choose app from the run/debug configurations drop-down menu.
3. In the target device drop-down menu, choose the AVD.
4. Go to Run.
5. Application will be installed on the AVD and emulator starts. Application starts running.

### Implementation

#### Module M1: Registration

This is onetime registration module where user can register once. In the registration user need to enter details like

First name  
Lastname  
Phone number

Password  
Address.

User get registered successfully after providing all the details. Only authentic users can use application (Fig. 2).

**M2: Admin Module** Admin module deals with administrator. Administrator will authenticate by providing basic details like.

Username  
Password

After authentication admin will be navigated to admin home page where there are buttons like add crop info, view reg users and logout (Fig. 3).

**M3: Add Crop Info** Authenticated admin can add crop information by providing details like.

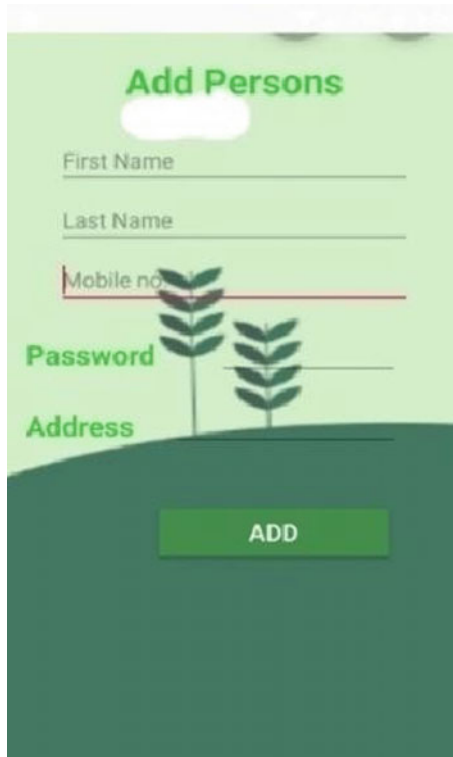


Fig. 2 Registration page



**Fig. 3** Admin page

Image  
District  
Year  
Season  
Crop name  
Soil type  
Area  
Production

Rainfall range (low, medium, high)

Crop details will be added to the server after providing all the inputs (Fig. 4).

#### **M4: View registered users**

Administrator has the privilege to view registered users. Details of the registered users appear with the following fields:

The image shows a web form for adding a crop. At the top is a teal button labeled "SELECT IMAGE". Below it are several dropdown menus: "Ananthapur", "2013", "Kharif", and "red". There is a text input field containing "Ragi". Below that is another dropdown menu with "red" selected. Further down are two more text input fields: "Area" with the value "2000" and "Production" with the value "3000". At the bottom is a text input field containing "higj". A teal "ADD" button is at the very bottom.

Fig. 4 Crop Addition

- First name
- Last name
- Mobile
- Address

See Fig. 5.

**M5: User Module**

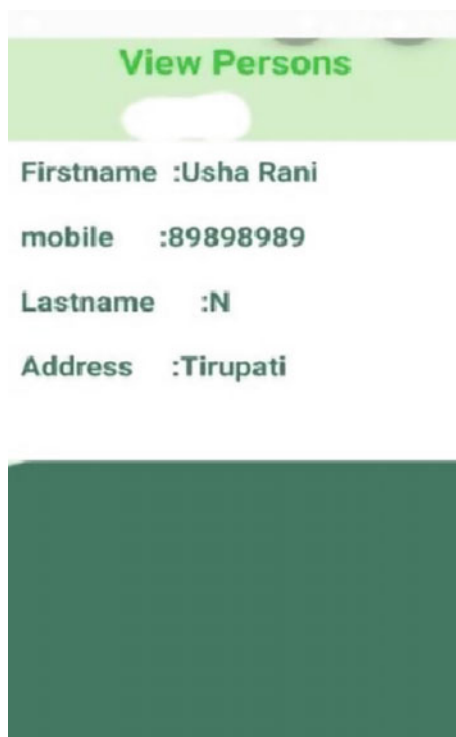
In user module authenticated user can login by providing details like username and password (Fig. 6).

User will be navigated to page where crop suggestions will be provided to the user by providing relevant details.

**M6: crop suggestion**

From the analysis of agriculture data crop suggestion will be provided to the user. For crop suggestion user need to provide the following inputs:

- State
- District
- Season
- Soil type



**Fig. 5** Persons viewed

Sources of irrigation

Rainfall range

Cost of cultivation

After providing inputs, output will be suitable crop, crop information and steps to grow that particular crop (Fig. 7).

## 4 Results

This work provides good suggestion for the farmer to choose better crop for the soil type, season and place. If sources of irrigation are there then farmer can get recommendations based on place, season and soil type, if there are no sources of irrigation then the farmer has to depend on rainfall based on the rainfall range (low, medium, high) farmer can get ideas. This application provides user friendly interface for the farmers to make better choice (Figs. 8, 9 and 10).

Fig. 6 Login page

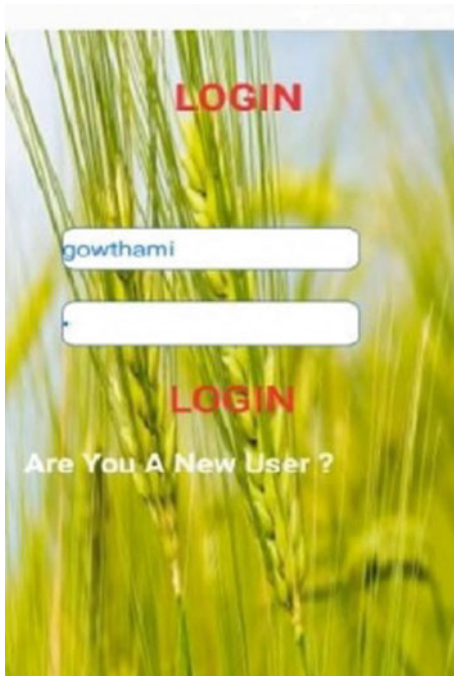
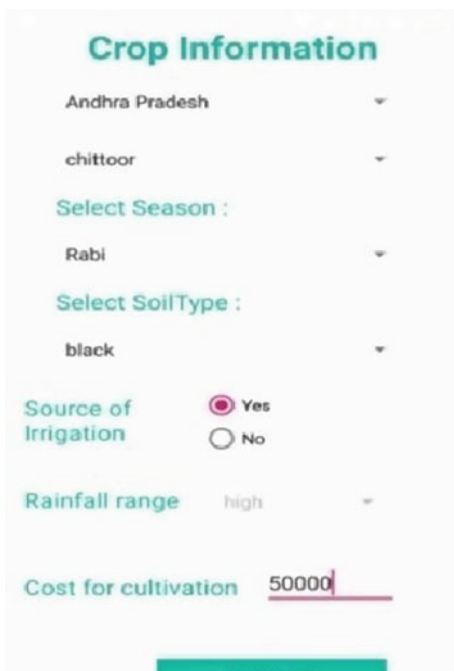


Fig. 7 Crop information





**Fig. 8** Crop list


## 5 Conclusion and Future Work

This SCS application is very helpful for the farmers to choose better crop from the analysis of agriculture data using parameters like state, district, season, soil type, rainfall range. This application is also helping for the farmers to increase productivity of the crops. The more productive the crops are the more benefit the farmer gets. This work can be further extended by providing videos to the farmer about cultivation of crops and adding the information about the profits farmer can get from particular crop. This work can also be extended by developing user interface in mother tongues. This work can also be extended by providing timely alerts to farmer about weather conditions and also providing information about modern techniques in agriculture.

**Fig. 9** Crop information



Ragi



**Crop Information** ▲

Finger millet is considered one of the most nutritious cereals. Finger millet contains about 5–8% protein, 1–2% ether extractives, 65–75% carbohydrates, 15–20% dietary fiber and 2.5–3.5% minerals. Of all the cereals and millets, finger millet has the highest amount of calcium (344mg%) and potassium (408mg%). The cereal has low fat content (1.3%) and contains mainly unsaturated fat. 100 grams of Finger millet has roughly on an

**Fig. 10** Steps to grow

## Steps to Grow



- Step 1. Use a crop calendar
- Step 2. Choose the best variety
- Step 3. Use high quality seed
- Step 4. Prepare and level the fields well
- Step 5: Plant on time
- Step 6: Weed early
- Step 7. Fertilize to maximize returns
- Step 8. Use water efficiently
- Step 9. Control pests and diseases effectively
- Step 10. Harvest on time

## References

1. <https://www.useoftechnology.com/technology-agriculture/>
2. <https://makeinbusiness.com/use-of-modern-technology-in-agriculture-an-overview-and-advantages/>
3. <https://www.downtoearth.org.in/blog/agriculture/applying-modern-tech-to-agriculture-66017>
4. R. Sridhar, K. Amol More, D. Ganesh Kenjale, A. Mahamadakram Desai, Kisan seva android application. *Int. Eng. Res. J. (IERJ) (Special Issue)*, 256–259 (2017). ISSN: 2395-1621
5. V. Patodkar, S. Simant, C.S. Shubham Sharma, S. Godse, E-agro android application. *Int. J. Eng. Res. Gen. Sci.* **3**, (3), (2015). ISSN: 2091-2730
6. M. Singhal, K. Verma, A. Shukla, Krishi Ville—android based solution for indian agriculture
7. A. Nirojan, V.N. Vithana, Smart cultivation partner mobile application (android) service to increase cultivation and sales. *Int. J. Sci. Res. Publ.* **7**(12), 111 (2017). ISSN: 2250-3153
8. S. Thankachan, S. Kirubakaran, E-agriculture information management system. *Int. J. Comput. Sci. Mobile Comput.* (2014)

# The Role of Long Non-Coding RNA (lncRNA) in Health Care Using Big Data Analytics



A. Revathi, S. Jyothi, and P. Swathi

**Abstract** In recent years the research in Bioinformatics is the major bigdata problem. In health care and Bioinformatics domain in order to protect the sensitive information, data sources publish only partial data. Analytics on partial data might be more complex and inefficient. The long non coding RNA's (lncRNA) are closely associated with human disorders and diseases like cancer, brain function and hereditary disease. So, in health care system to predict the diseases caused by long non coding RNAs is a challenging task. Analysing lncRNA from large number of RNA sequences is a difficult task to be solved. There are so many computational methods used for identification, classification of lncRNA and its functionalities. Recently, Machine learning and deep learning has been employed for ncRNAs identification and classification and has shown promising results. So, aim of this study is to specify the analytics performed in all the approaches to know the role of lncRNA.

**Keywords** Bioinformatics · Genomics · RNA · Non-coding RNA · Long non-coding RNA · Big data analytics · Health care system · Machine and deep learning · Cloud data analytics

## 1 Introduction

In Health care system the volume of data is growing fast because of the data contains of medical images, clinical reports, medical sensors, genome sequences, experimental studies of biomedical research, gene expression profiles and so on [1]. In

---

A. Revathi (✉) · P. Swathi

Research Scholar, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [arevathi20@gmail.com](mailto:arevathi20@gmail.com)

P. Swathi

e-mail: [swathivinubaby@gmail.com](mailto:swathivinubaby@gmail.com)

S. Jyothi

Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020

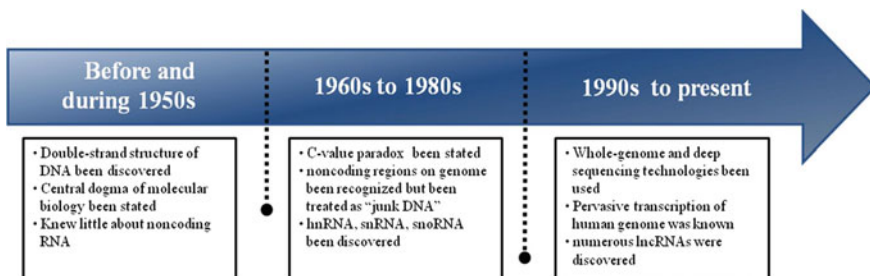
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_35](https://doi.org/10.1007/978-3-030-46939-9_35)

bioinformatics research, the size of sequence in genome of human is nearly 200 gigabytes and 20 petabytes of genomic data are maintained by European Bioinformatics Institute (EBI), which is the biggest data repositories of biology, while the PubMed stores about 24.6 million records of biomedical literature in 2014, in 2020 it is expected to 25,000 petabytes of data will be generated in health care domain [1]. The EBI increases the computing power every month due to rapid increase of data in bioinformatics [2]. The national center of biotechnology information (NCBI), USA is maintaining the huge biology databases and distributing it throughout the world. With the help of high volume of data in bioinformatics accurate analytics can be done [2]. Present days the biologist not depending on labs to find a new disease biomarker, now they depend on genomic data which is available with help of big data technologies with lesser cost most effective such as automated genome sequencers [2].

Next generation sequencing (NGS) or High-Throughput Sequencing (HTS) techniques are used to handle a huge number of DNA or RNA molecules quickly and collectively [3]. These NGS technologies, such as whole-genome sequencing (WGS), whole-exome sequencing (WES), and/or targeted sequencing, are gradually more applied to the study of biomedical and medical practice to find the disease and/or drug-associated with variations of genetic for the advance precision medicine [2] (Fig. 1).

It found that the transcribed genome consists only 2% encodes protein genes and the remaining 98% are noncoding transcripts [5]. From recent years this non coding transcripts drawn major attention. The transcripts which are longer than 200 nucleotides are long non-coding RNAs (lncRNAs), which are more in number and plays main role in various regulatory levels such as transcriptional regulation and post transcriptional regulation. The non-coding RNA (ncRNA's) is functional RNA, which is incapable of being translated into protein [3].

At the beginning the non-coding RNA is treated as garbage data, later researches found some biological functions are associated with ncRNA which cause diseases like cancer, brain function, Alzheimer and other hereditary disease. The biological experiments and computational methods found many functionalities of Non coding



**Fig. 1** History of discover of long non-coding RNAs [4]

RNAs (ncRNAs). In the early stage the biological experiments used the technologies such as genomic tiling arrays in the transcriptomes of organisms and full-length complementary DNA cloning. This experiment found non coding RNAs inefficient way, but more costly because they always require a greater number of RNA samples, to overcome this problem researchers developed the computational Biological approaches to identify ncRNA. They combined computational approaches with experimental methods [3]. Later the next generation sequencing technologies (also known as deep sequence technologies) are developed and discovered long ncRNAs and small ncRNAs. This technology reduced the cost when compared with conventional sequencing technologies. New transcripts are identified with the help of next generation sequencing techniques that made the desire to develop computational methods to identify non coding RNAs effectively and efficiently [3].

Large amount of sequence data produced by the genomics know as big data that is very difficult process and manage on normal machines. Big data technologies become solution for this problem by providing the platform for more storage and with high computational power. The analytics of big data in bioinformatics required to be well addressed from the perspectives of big data technologies and current data analytic approach because only 2% encode proteins found from 3 billion base pairs (bp) of the human genome, the rest is NON-coding (ncRNAs) in which only few functionalities are known, the remains are unknown [6].

Due to vast amounts of data in human next generation sequence (NGS), the structure and function of non-coding RNA analysis is a difficult task. For many years in Bioinformatics the main research area is classification of non-coding RNA's [7]. Based on length of RNA, the ncRNAs are of two types small non coding RNAs length is shorter than 200 nucleotides and long non-coding RNA'S length is longer than 200 nucleotides There are many non-coding RNA's types such as long non-coding RNA's (lncRNA), small RNA's, micro RNA's, transfer RNA's (tRNA's) and ribosomal RNA (rRNA) etc. The classification helps to identify the different biological functions associated with different types of RNA'S. So, classification of ncRNA's is very important in identifying the functionalities [7] (Fig. 2).

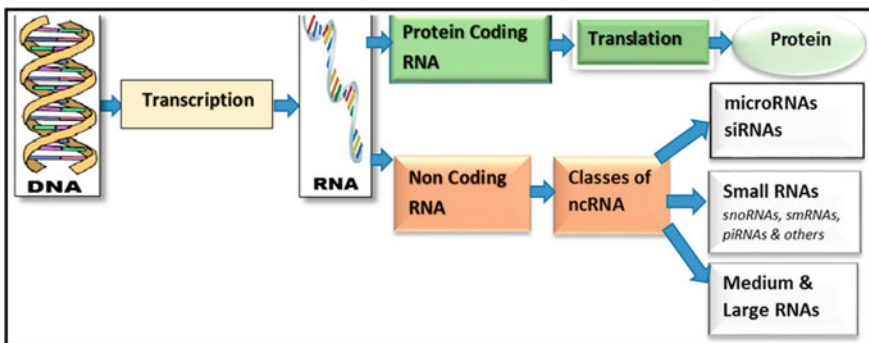


Fig. 2 Coding and non-coding RNA genes transcription process for protein production [3]

The main focus of this paper is to describe all the available analytics, databases and tools of the long non-coding RNAs in Health care system. The remaining paper consists of Sect. 2: descriptive outline of the big data analytics methods used to identify, analyse and classify of lncRNA. In Sect. 3: various machine learning and deep learning techniques are described and Sect. 4: specifies the methods of cloud data analytics to handle massive amounts of lncRNA. Thereafter, all the methods are described in tabular format, finally conclusion and future scope of research are provided.

## 2 The Role of lncRNA in Health Care Using Big Data Analytics

Big data V'S of 'Big' biological data are defined with 5 V's. The *Volume* indicates the huge amount of data generated due to patient's clinical data, experimental research studies and genomics. The *Variety* is of data from different sources of health care. *Velocity* is the speed at which the data is generated from real time scenarios. *Veracity* indicates the degree of inconsistent biological data generated. *Value* is the ability to take decisions by understanding the managed biological data.

The big data has made very great booming effect on Bioinformatics from the past few years [6]. To understand the field of Bioinformatics the researchers from all over the world as made several attempts by analysing and using big data tools [6]. For many years the research is done in genes and gene sequencing, the new outlooks on the human genome and the progress made in big data analytics attracted the researchers [7]. The Human Genome Project is the highest achievement accomplished in the early 2000s. The biologist says that in 2025, the single genome stored data are 30 times more than the size of the genome itself, that can sequence 100 million to 2 billion human genomes. The more supplies are required for serious study of genetics work [6]. To handle the speed of analysis and volume of data major changes has to be done. The big data analytics helps the scientist to look more closely to human genes and have much progress in the issues of it [7].

NGS is a power full tool for genomics research, that produce vast volumes of data. NGS analytics with big data applications of Hadoop based are supporting users with huge volumes of datasets in a distributed manner which is parallelized. On single machine NGS platforms can read millions of sequencing in parallel, which produce a signal data of many terabytes, corresponding to the sequence data of many hundreds of gigabytes [8]. NGS analytics can analyse very large data sets in shorter time span with the help of big data technologies that provide filtering and mapping [8]. The present tools for processing big data are accurate and faster for analysing the huge input sequence data to identify the patterns in sequencing which are hidden, such as regulatory monitoring, interactions of the bio molecules and so on [8].

From past many decades, thousands of lncRNAs have been identified which are related to the development of diseases like epigenetics, cancer, brain function and

hereditary and so on [9]. RNA roles are identified that include the synthesis of protein, processing of nucleic acid and gene regulation. Recent studies in humans have shown that protein coding genes are a small fraction of the genome, while the non-coding RNAs represent the huge majority (i.e. more than 80% in mammals) and according to the recent studies 15,000 long non coding RNAs in humans are estimated, it is very difficult to identify the lncRNAs expression profiles specific to cell, cell cycle tissue and developmental stage of disease [9].

The Long non-coding RNA (lncRNA) play a main role of various diseases occurrence and development, it is very difficult to detect the disease caused by the lncRNA from the huge amount of biological data, so computational methods are required to find it [10]. Many human diseases are associated with lncRNAs [11]. According the lncRNA Disease database 150 human diseases are found due to the lncRNAs [12], such as breast cancer [13, 14], leukemia [15, 16], colon cancer [17], prostate cancer [18] and Alzheimer's disease [18]. The evidences say that the lncRNA is a potential biomarker of human disease and drug target in drug discovery and clinical treatment, so it very important and urgently required to identify the potential lncRNA associated diseases.

The big data problems are different from the challenges of big data. The first problem in Bioinformatics is that the data is increasing in terms of dimension and all the over the world the number of instances is distributed. only the part of these data transferred over the internet, the remaining are not transferred due to privacy and ethical issues. The second problem is Bioinformatics data is heterogeneous in nature, which leads to so many analytic problems. The Bioinformatics existing tools are not suitable for big data. Using Hadoop and Map reduce platform the tools Bio Pig and crossbow are developed for sequence analysis.

lncRNA identification is a challenging task, even though there are several computational methods which are already developed for lncRNA identification from other RNAs [19].

## ***2.1 The list of tools based on sequence features are [19]***

1. Coding potential calculator (CPC)
2. Coding potential assessment tool (CPAT)
3. Coding non coding index (CNCI)
4. Predictor of long non coding RNAs and messenger RNAs based on an improved k-mer scheme (PLEK)
5. Long noncoding RNA IDentification (lncRNA-ID)
6. lncRNApred
7. lncRScan-SVM.



## 2.2 The List of Tools Based on Structure Features [19]

1. lncRNA-MFDL
2. RNAfold.

## 2.3 lncRNA Databases

In the public data huge lncRNAs available, but it is very limited for the literature supported evidence and biological research activities [20] (Table 1).

The lncRNAs databases consist of 2000–70,000 transcripts. Non-code v3.0 is the largest database which stores 73,000 transcripts and more than that [21]. These databases automatically generate a list or a table of visualizations of query results and also helps in graphical visualizations such as plots. The majority of the databases, can be used to download entire or portion of their data as files. Those files can be used for development of new computing tools or can be used for further analyses [21]. The originating source of the transcripts is not known for the available databases due to insufficient information [21].

To identify the genes from large sequences of ncRNAs, the functional annotation plays a vital role. The role of long ncRNAs (lncRNAs) which are not involved to produce protein-coding has gained lots of interest in recent years [22]. At present days by using the results of the functional annotation of ncRNAs or lncRNAs is the hot research topic in the domain of next-generation sequencers (NGS) [22]. The lncRNAs plays a main role for development of many diseases by regulating the gene expression and epigenetics. To know the function of lncRNAs, many studies have been conducted [22].

**Table 1** Available lncRNA databases [19]

Database	Description
CHIPBase	The transcriptional regulation of long noncoding RNA database for decoding a
DIANA-LncBase	The targets of long noncoding RNAs are verified experimentally and computationally predicted
Lncipedia	LncRNA transcript sequences and structures for annotated human database
LncRNAdb	The eukaryotic long non coding RNAs annotations are provide in this database
LncRNADisease	The lncRNAs disease associations are supported experimentally
LncRNome	Humans long non coding RNAs comprehensive database
Noncode v3.0 The functional lncRNA database	It contains human RNAs of annotated protein-coding

**Table 2** The computational methods and tools for functional annotation of non-coding RNAs [23]

Computational methods	Features	Tools
Hidden Markov model (HMM) + Covariance model (CM)	ncRNAs annotation is faster	CM-HMM
Support vector machine (SVM)	lncRNA genes are classified and identified	DREME
Ensemble classifiers	The genomic variant Ensemble classifiers are annotated	VEP
Deep learning neural network (DL-NN)	Cis-regularity regions are identified	DECRES
Support vector machine (SVM)	lncRNAs are identified with several features	LncRScan-SVM
Deep mining algorithm (DMA)	lncRNAs are identified by deep mining algorithm	DM lncRNAs
k-mer scheme and SVM	lncRNAs and mRNAs are differentiated	PLEK
Random forest Algorithm (RFA)	It has multiple features	LncRNA-ID

The computational methods depend more on known proteins for classification of ncRNAs, which makes it difficult to differentiate coding from non-coding transcripts. For identification of protein-coding, the recent approaches have only used SVM, HMM, NN or statistical models to get 90% of accuracy [23]. These computational methods are listed in Table 2. To resolve the above issues, there is a need for development of an advanced computational methods [23].

### 3 The Role of lncRNA in Health Care Using Machine and Deep Learning Analytics

The statistical tests of control and experimental conditions are not suited for analysing the currently available data. Machine learning strategies are now available to fill the gap. Machine learning algorithms are of two types unsupervised and supervised algorithms. Both algorithms are required and best suited to focus effectively on “big data” biomedical problems. The effective integrative analyses can be performed by machine learning methods to resolve the issues of big data in biology [24]. For descriptive and predictive analytics of lncRNAs databases, the commonly used tools of machine learning approaches are supervised, unsupervised and hybrid methods.

Machine learning models are used to predict lncRNA disease associations. But these models have their own advantages and disadvantages [24]. To understand human complex diseases at lncRNA level require an effective computational model that can use heterogeneous biological data generated from different data sources that could benefit more effective identification of new lncRNA disease interactions [24].

The conservation and development of tissue specific expression is low are lncRNAs due to this the functional annotation is very difficult. Throughout the biological processes the lncRNAs are involved [25]. The subcellular localization gives the most information regarding the biological function of lncRNA, but it is very difficult to discover it despite some prediction methods are present. With the help of deep learning algorithms, a model called DeepLncRNA developed to predict the lncRNA subcellular localization from lncRNA transcript sequences [25]. This DeepLncRNA is able not identifying the disease associated point mutations. The functional annotation of lncRNAs which is identified by DeepLncRNA can be used for future research [25] (Table 3).

## 4 The Role of lncRNA in Health Care Using Cloud Computing

The rapid increase of various healthcare devices and applications which are generating varieties of data, made very difficult for health care organizations to analyse the large-scale data [25]. To take a better decision, the data need to effectively analysed and processed. The traditional systems are not having ability to process the huge amount data generated by the healthcare system. So, cloud, environment with the distributed system is required to solve the scalability issues [25]. The cloud technology provides on-demand services for analysing, processing and storage. To process large scale data sets on distributed environment Hadoop can be deployed on cloud environment. Instead of using traditional software for healthcare applications, it is better to have real time information with the help of cloud services to have quality healthcare [25].

### 4.1 Cloud Platforms for Genomics [20]

S. No	Name	Description
1	Google Genomics	<ul style="list-style-type: none"> <li>• The data produced by the genomics is stored, processed, explored and shared by this cloud platform</li> <li>• It uses MapReduce for parallel computing to process data in minutes or hours</li> </ul>
2	DNAnexus	<ul style="list-style-type: none"> <li>• Large growing sequence data is handled by using this platform</li> <li>• Various bioinformatics tools for sequencing can be accessed</li> <li>• Genomic data can be visualized quickly</li> <li>• Provides fast visualization of genomic data</li> </ul>

(continued)

(continued)

S. No	Name	Description
3	Globus Genomics	<ul style="list-style-type: none"> <li>It is graphical workflow environment with good computing infrastructure that provides algorithms, data management tools that are easy to use</li> </ul>

Now a days the big biomedical data is analysed by cloud computing which is offering massive scalable computing and storage, with secure data access, and the resources and applications are provided on demand access [19] (Fig. 3).

A health care framework for data processing in cloud environments that runs Hadoop clusters can be used to run applications that can fit a large number of users and used to integrate data together using Hadoop MapReduce to get results quickly. prevention of various diseases can be possible with the big data analysis of health care on a cloud that helps to reduce the cost of healthcare [19] (Fig. 4).

The Bioinformatics, cloud-based services are classified into Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) in order to address the big data storage and analysis issues. The supercomputer's computational power is now easily accessible on available demand at low cost with cloud computing. AWS Lambda, Google Cloud Functions, Microsoft Azure Functions and IBM Cloud Functions are the major public cloud vendors which provide FaaS computing platform [27]. The Function-as-a-service (FaaS) platform where available 24/7 to provide resource provisioning, monitoring, scaling and fault tolerance are provided automatically [27]. This serverless FaaS platform can also easily de ploy the code snippets to the cloud in the form of microservices.

The Bioinformatics tools are not having cloud computing support, as they are developed for desktop applications, due to this Bioinformatics task in the cloud are complex. The popular RNA-seq mapper with high speed it performs highly accurate spliced alignment, but is not developed for the cloud environment [19]. Huge datasets can be processed quickly and safely by MapReduce and Hadoop allows distributed processing of large data sets with commodity hardware is the advancement of cloud. The traditional Bioinformatics tools and algorithms are redesigned for Hadoop MapReduce infrastructure to have the support and the benefits of it [19]. The main purpose of Hadoop frame work is to distribute data and it is processed by the Apache Spark frame work. A new tool SparkBWA (Burrows-Wheeler Aligner) is developed to boost the performance of sequence aligner. Sparkseq was created to study the utility of Apache spark in the genomic context [19].

The AWS delivers all the public datasets, as services and centralized data repository. The Amazon EC2 OR Amazon EMR cluster is very useful for mapping the human genome and to locate, download, customize, and analyse, but in the past it takes hours or days to do all this [19].

**Table 3** The deep learning approaches for lncRNA classification [20]

Model name	Training procedure	Description
Auto Encoder (2017)	Like other neural network such as an SVM is used to perform the actual classification by AEs are trained until the desired output by using backpropagation. Then it can be used to train a simpler model	The aim of feed forwarded neural network is to reproduce the input as output. The encode compresses the input to a lower-dimensional representation and then output is decompressing back. The output dimensionality matches exactly, but differs from the input due to compression. The latent space becomes a useful representation of the input, due to AE minimizes the error during training
Deep belief network (2018)	The training is made shallow to optimize by applying a greedy learning approach form top down manner to one layer at a time	The training is made shallow to optimize by applying a greedy learning approach form top down manner to one layer at a time
RNN (2018)	The training of RNN is time consuming and difficult due to replication of units for each step of backpropagation	RNNs are suitable for sequential data due to persistent memory and the neural network contains loop to feedback one layer by itself
CNN (2018)	To minimize the loss between the actual output and desired output, the backpropagation training is applied	To model the neighbouring pixels the CNN uses spatial convolution for specialized image data. The pooling reduces the image dimensionality due to set of filters that extracts spatial features. Usage of same filter for each location of the image is useful in sparse two dimensional data like matrices and images

## 5 Limitations and Future Scope of Research

The present tools and computational methods cannot be retrained or tailored by users and they are not customized to meet the requirements of researchers. The databases and other platforms are not sufficient for analysis, the lncRNA are lacking of experimental data. The major problem is lack of standardization to explore the efficient existing lncRNAs data. In spite of knowing the importance of lncRNAs, the molecular mechanisms and functions of it is not yet understood fully which is a one of the challenge areas to work on lncRNAs [28].

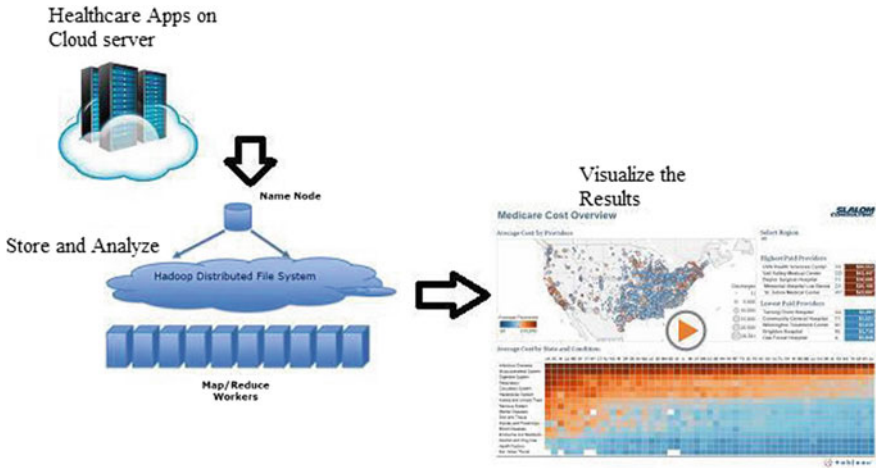


Fig. 3 Framework for healthcare data processing based on Hadoop [26]

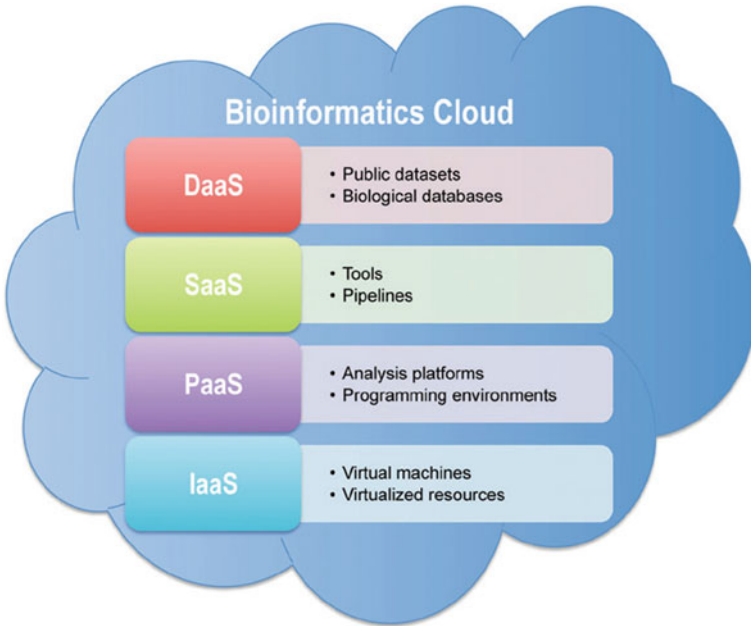


Fig. 4 Bioinformatics cloud-based services [18]

The NGS tools are not supporting cloud environment due to vast amounts of NGS data and bottleneck the analysis of data remains as a challenge, so in the future an open Bioinformatics cloud to be built to have benefits of cloud services. To analyse the RNA-seq data a web application of cloud computing is required for analysing the complete workflow. There is requirement of Apache spark-based bioinformatics algorithms for genomic data analysis. The protein subcellular localization prediction is a well-established research field, but prediction of the lncRNA localization need to fill research gaps.

## 6 Conclusion

The pre and post classification are not required by the machine learning and deep learning algorithms to handle the big genomics human data. To improve present tools and approaches a careful attention is required and need to specify future directions to overcome the various limitations of available resources for lncRNA [29].

## References

1. H. Kashyap, H.A. Ahmed, N. Hoque, S. Roy, D.K. Bhattacharyya, Big data analytics in bioinformatics: "A machine learning perspective. Journal of latex class files" big data analytics for genomic medicine. *Int. J. Mol. Sci.* **18**, 412 (2017)
2. N. Amin, A. McGrath, Y.P. Chen, Evaluation of deep learning in non-coding RNA classification. *Nat. Mach. Intell.* **246**(1), 246–256 (2019)
3. Q. Abbas, S.M. Raza, A.A. Biyabani, M.A. Jaffar, A review of computational methods for finding non-coding RNA genes. *Genes* **7**, 113 (2016)
4. M.-H. Bao, V. Szeto, B.B. Yang, S. Zhu, H.-S. Sun, Z.-P. Feng, Long non-coding RNAs in ischemic stroke. *Cell Death Dis.* **9**(3) (2018)
5. X.N. Fan, S.W. Zhang, lncRNA-MFDL: Identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Mol. BioSyst.* (2015)
6. R. Martiz, M.A. Supaksha, N. Hemalatha, Application of big data in bioinformatics a survey, in *International Journal of Latest Trends in Engineering and Technology Special Issue SACAIM*, pp. 206–212 (2016)
7. I.V. Novikova, S.P. Hennelly, C.S. Tung, K.Y. Sanbonmatsu, Rise of the RNA Machines: Exploring the Structure of Long Non-Coding RNAs. (Elsevier, Amsterdam, 2013)
8. Z.H. Guo, Z.H. You, Y.B. Wang, H.C. Yi, Z.H. Chen, A learning-based method for lncRNA-disease association identification combing similarity information and rotation forest. *IScience* **19**, 786–795, (2019)
9. R. Tripathi, P. Sharma, P. Chakraborty, P.K. Varadwaj, Next-generation sequencing revolution through big data analytics. *Front. Life Sci.* **9**(2), 119–149 (2016)
10. H. Kashyap, H.A. Ahmed, N. Hoque, S. Roy, D.K. Bhattacharyya, Big data analytics in bioinformatics: a machine learning perspective. *J. Latex Class Files* **13**(9) (2014)
11. G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, Q. Cui, lncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **41**(Database issue) (2013)

12. R.A. Gupta, N. Shah, K.C. Wang, J. Kim, H.M. Horlings, D.J. Wong, M.C. Tsai, T. Hung, P. Argani, J.L. Rinn, Y. Wang, Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**(7291), 1071–1076 (2010)
13. A. Guffanti, M. Iacono, P. Pelucchi, N. Kim, G. Soldà, L.J. Croft, R.J. Taft, E. Rizzi, M. Askarian-Amiri, R.J. Bonnal, M. Callari, A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics* **10**, 163(2009)
14. G.A. Calin, C.G. Liu, M. Ferracin, T. Hyslop, R. Spizzo, C. Sevignani, M. Fabbri, A. Cimmino, E.J. Lee, S.E. Wojcik, M. Shimizu, Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas (2007). <https://doi.org/10.1016/j.ccrjuly>
15. L. Pibouin, J. Villaudy, D. Ferbus, M. Muleris, M.T. Prospéri, Y. Remvikos, G. Goubin, Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. *Cancer Genet. Cytogenet.* **133**, 55–60 (2002)
16. S. Chung, H. Nakagawa, M. Uemura, L. Piao, K. Ashikawa, N. Hosono, R. Takata, S. Akamatsu, T. Kawaguchi, T. Morizono, T. Tsunoda, Y. Daigo, K. Matsuda, N. Kamatani, Y. Nakamura, Y. Kubo, Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.* **102**(1), 245–52 (2010). ISBN: 1349-7006.2010.01737.x
17. M.A. Faghihi, F. Modarresi, A.M. Khalil, D.E. Wood, B.G. Sahagan, T.E. Morgan, C.E. Finch, G.S. Laurent III, P.J. Kenny, C. Wahlestedt, Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of  $\beta$ -secretase expression. *Nat. Med.* **14**(7), 723–730 (2008)
18. L. Dai, X. Gao, Y. Guo, J. Xiao, Z. Zhang, Bioinformatics clouds for big data manipulation. *Daietal. Biol. Dir.* (2012)
19. S. Fritah, S.P. Niclou, F. Azuaje, Databases for lncRNAs: A Comparative Evaluation of Emerging Tools (Cold Spring Harbor Laboratory Press, New York, 2019)
20. N. Amin, A. McGrath, Y.P. Chen, Evaluation of deep learning in non-coding RNA classification. *Nat. Mach. Intell.* **246**(May), 246–256 (2019)
21. C.M. Sreeshma, M. Manu, G. Gopa Kumar, Identification of long non-coding RNA from inherent features using machine learning techniques, in *IEEE* (2018). ISBN: 978-1-5386-6434-6©
22. Q. Abbas, Current challenges of computational intelligent techniques for functional annotation of ncRNA genes. *Int. J. Med. Res. Health Sci.* **8**(6), 54–63 (2019)
23. S. Kaur, S. Kaur, Genomics with cloud computing. *Int. J. Sci. Technol. Res.* **4**(04) (2015)
24. X. Chen, C.C. Yan, X. Zhang, Z.H. You, *Long Non-Coding RNAs and Complex Diseases: from Experimental Results to Computational Models.* ( Oxford University Press, Oxford, 2016)
25. B.L. Gudenias, L. Wang, Prediction of lncRNA Subcellular Localization with Deep Learning from Sequence Features. Published in scientific reports (2018)
26. S. Rallapalli, R.R. Gondkar, U.P. Ketavarapu, Impact of processing and analyzing healthcare big data on cloud computing environment by implementing Hadoop cluster. *Procedia Comput. Sci.* **85**, 16–22 (2016)
27. M.X. Liu, X. Chen, G. Chen, Q.H. Cui, G.Y. Yan, A computational framework to infer human disease-associated long noncoding RNAs. *PLoS One* **9**(1), e84408 (2014)
28. L.H. Hung, D. Kumanov, X. Niu, W. Lloyd, K.Y. Yeung, Rapid RNA sequencing data analysis using serverless computing. CC-BY-NC-ND 4.0 International license, 13 Mar (2019)
29. Y. Fang, M.J. Fullwood, Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics Proteomics Bioinforma.* **14**, 42–54 (2016)



# A Framework for Modeling and Analysing Big Biological Sequences



Sai Jyothi Bolla and S. Jyothi

**Abstract** Increasing volumes and variety of biological data makes analysis more challenging. Storing those volumes of heterogeneous data is a bigger challenge. Biological Sequence analysis is more important in bioinformatics which affects the healthcare system. Genomic sequence analysis can predict the risk of a disease before it becomes a problem. These analytics can improve the overall patient's health, improve patient's experience and reduce the cost of healthcare system. Towards the end of the paper, a framework is proposed which can effectively model and analyse biological sequences in parallel distributed computing environment using NoSQL databases.

**Keywords** Biological sequence analysis · Bioinformatics · Genomic sequence · Healthcare system · NoSQL databases

## 1 Introduction

Data Modeling is a process of ordering the data in a specific method to an application. More technically, Data modeling is way of envisaging how data is related to each other and how it is analyzed and saved in a system. Indeed, Data Modeling helps in understanding the behaviour of the real world data.

---

S. J. Bolla (✉)

Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, India  
e-mail: [saidilpyerram@gmail.com](mailto:saidilpyerram@gmail.com)

S. Jyothi

Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_36](https://doi.org/10.1007/978-3-030-46939-9_36)

## ***1.1 Merits of Data Modeling***

**Managing Data as a Resource** Data modeling gives way for the data for normalizing and defining it in terms of as it is and what attributes it can take. To query the Data Base and retrieve reports out of it, there are several tools in Data Modeling. However, having a good data model may enable one to have a great amount of data but it is just not enough to take benefit out of it.

A decent data model along with a decently designed database, end users can access information which—they hardly realize of its collection for analysis.

**To Integrate Existing Information Systems** Often the vendors end up in the position with data in a various systems that does not communicate internally. By way of modeling the data in each of these systems, relationships are defined and redundancies are identified to resolve the discrepancies and unite the disparate systems so as to make them communicate with each other.

**Designing Databases and such Repositories** Modeling data is crucial in devising a decently working database. Data modeling gives way for making better decisions about data warehousing and repositories. With a clear view of the data in possession, one can disclose if there is a need for a warehouse at a global level, or just an independent data mart, or else a series of interlinked data marts. Modeling data can also aid in deciding if there is a need for a relational database or a NoSQL database. Modeling data is the most effective way to infer the various needs in a business pertaining to data storage and service.

**Understanding the Business** Data modeling is way to comprehend the typical challenges in a business in order to finalize the data that drives it. In the process of building a database pertaining to customers, as to illustrate, one needs to understand what kind of customers' data is sampled and how is it utilized. The data and relationships present in the data model provide a basis on which one need to build a comprehensive approach of business processes.

**Business Intelligence** Utilizing a proper modeling and the subsequent reporting, one can spot the trends in the business, expenditure patterns so as to make predictions that help resolve challenges and find the opportunities in a business.

**Knowledge Transfer** Data modeling is but a kind of evolving a document benefits both the business personnel and the technical personnel to use it. The document provides a common vocabulary that various job roles share, and to provide the newcomers business glossary to convey enhanced information about the business. As a helping hand, a dictionary of data built out of a well executed data modeling is indispensable [1].

## 2 Big Data and Modeling

Conventional tools of database management along with traditional data processing apps fall redundant with regard to processing the big data which consists of large and complex data sets. The method of analyzing large amounts of such data to unravel the dormant patterns and secret correlations is therefore termed as big data analytics. Big data may generate from sensors and social networking web sites such as Facebook and Twitter. Big Data is well known for its ability to deal with the data with the 5 characteristics such as volume, variety (structured and unstructured data), velocity (high rate of changing), veracity (biases, noise and abnormality), validity (correctness and accuracy of data), volatility (how long data is valid and how long it should be stored), value (Sources the value of the data to releases the dormant knowledge in it for those who can deal with the large scale data) and visualization. However, the conventional definition omits two very important features which distinguish big data from the traditional databases and data warehouses. Firstly, big data as such allows increase in data from time to time as new data gets added in a dynamic way to the pool of already existing content. Secondly, big data are geographically distributed. The sources of Big data went beyond the experiments in particle physics or logs and indexes of search engines.

Currently, the Internet usage across the globe contributes to adding of the volume of data as big as terabytes and petabytes. As a result, the volume of data so added up gives way for typical data generated by different applications making it richer day-in day-out. Hence, the conventional relational databases face a challenge to capture, store, search, share, analyze and visualize the data. However, many IT companies face the challenge of managing the big data either using a NoSQL (“not only SQL”) database, such as Cassandra or HBase, or they may resort to a distributed computing system such as Hadoop [2]. As NoSQL databases host key-value stores which are non-relational, distributed across, and horizontally scalable while schema free, the question arises, whether do we need Data modeling as on date? Conventional data modeling methods concentrate on determining the intricate relationships among schema-enabled data. Nevertheless, these methods fall short to determine non-relational, schema-less databases. Hence, conventional ways of data modeling renders obsolete. Hence a novel technology to manage big data for optimum business value is the need of the hour. Big Data modeling becomes significant to map all these varieties of data as it consists of all the structured, semi-structured and unstructured data. Therefore, Big Data must be modeled prior to applying data analytics.

### 2.1 Big Data Model

Big data model is otherwise an abstract layer envisaged to process the data in various physical devices. Times we are in give way for volumes of data in various formats saved across the globe in various servers. Big data model gives an opportunity to

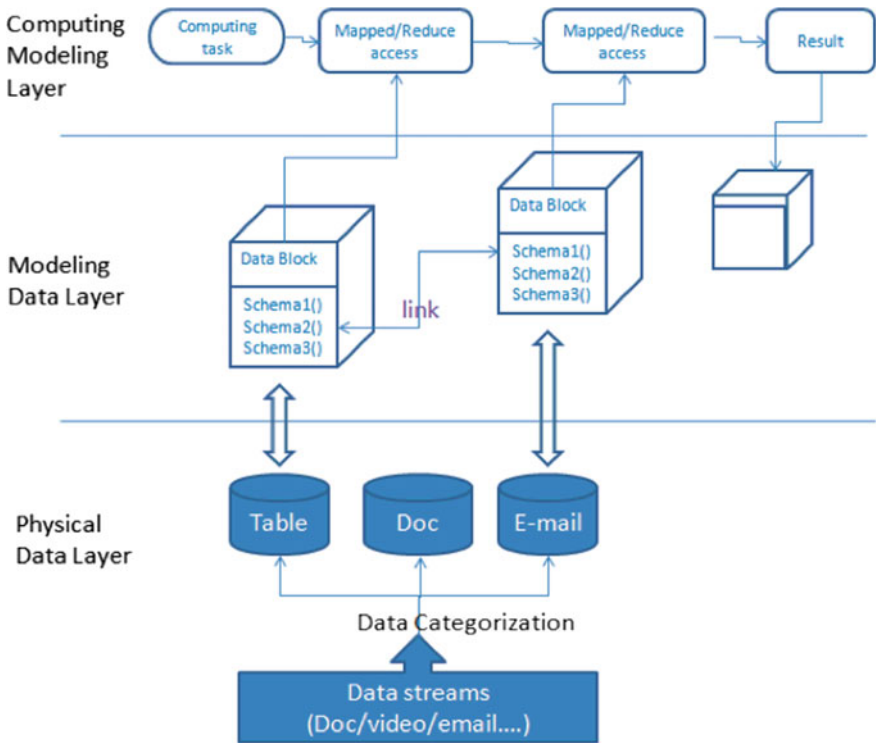


Fig. 1 Big data architecture

process various data resources creating a fundamental architecture of data so as to have more optimized applications using cost effective computing tools. Figure 1 shows the architecture of Big data in general [3].

Diagram above shows the three model layer architecture. Physical data layer in the big data system hosts various sorts of data like audio, video, business tables, log reports generated by various gadgets and so on. An abstract data model is built to process the physical data as a part of abstract data model. Subsequently, the application layer also called as modeling layer is used to retrieve business value information. With the three layered architecture, data models are built to distinguish physical data and data use. Therefore, the built application accesses the data through the data model instead of accessing the physical data. This method enables the application flexible and even manages the data.

To begin with, one has to create data blocks basing on data storage, data type, relationship, read-write requirement and so on to construct a big data model. Also, we need to have modeling apps maintained in these models, so as to enable the data models to display and save the current data.

### 3 Bioinformatics and Big Data

Bioinformatics is an interdisciplinary field that develops software tools and techniques for gathering, storing, analyzing and integrating biological information. It is a known fact that the data volume grows in bioinformatics research at instant speed. Owing to the vast diversity and its scale new technics and algorithms are demanded to deal with. Besides, new analytics are required to extract the value as well the dormant knowledge in it.

The field of bioinformatics provides tools to analyses facilitating understanding of the life's molecular mechanisms, analyzing and correlating genomic and proteomic information at large. Massive amounts of genomic information, including both genome sequences and expressed gene sequences, being available, more efficient, sophisticated and even specific analyses became essential.

#### 3.1 *Biological Data Characteristics*

Biological data has specific features that renders the task of managing biological data challenging.

1. Biological data is highly intricate in comparison to most of the other data. The very aim of drawing inferences from biological data is to present intricate substructure of the data besides the relationships so as to make sure that total information during the modeling of biological data is arrived at. The data model is supposed to present any level of intricacy in a data schema, relationship, or schema substructure in a hierarchical, binary, or tabular data format. For instance, a biological sequence is treated as a simple integer coordinate system by the NCBI biological data model which gives scope for diverse data. A varied amount of data is intimately connected to the coordinate system like that of sequences in amino acids.
2. As variability range in the data is high, it requires a agility in operating the data types and values. However, the frequency in data exception in the biological data structures seeks an option of types of data for a given chunk of data. However, overlapping is found often in the data types among the various organisms and genome projects.
3. In biological databases, Schemas vary at a swift pace. Hence, it is not possible in most relational and object database systems as of now to expand the schema.
4. Different biologists representing the same data would mostly be different (even if the same system is used).
5. Most of the biologists do not care to know about the design of schema or even the data structure. Thus, the interface of the biological database/resource must be able to display the information to the end user in a suitable way as to address the problem and to reflect on the underlying data structures.

6. In Biological applications, the context plays a crucial role in appending meaning for its use in a more meaningful way. Contexts integrated more in number provide a rich understanding of a biological data value; whereas the isolated data values are redundant in biological systems. Much in a similar way, “671” is pointless without units and context. Eg. gene 671, gene starts at position 671, gene culminates at a position 671, with taxonomy identification 671, number of bases 671, etc.
7. For a biologist, designing and presenting complex queries are far more important. Complex queries must be available for him even without knowledge of the data structure. To construct complex queries for an end user it is very difficult and so a tool for building these queries is highly appreciable if it is with some predefined query templates.
8. Past versions of existing data provide crucial information to the users of biological information and hence they opt for access to the old records. For example, GenBank utilizes the Accession along with the version number for protein sequences in the flat file records; the version is commensurate with the gi number represents a specific sequence. While the accession is unaltered during the updates, protein sequence however, receives a new gi during updating. GenBank is otherwise a nucleotide-centric public database, wherein proteins are but the translated products of nucleotides [4, 5].

Owing to these typical traits, current DBMS are not competent to solve completely the complex biological data giving in for more scope for research in database management. Besides, the traditional modeling and computing techniques run absolute to process the data in bioinformatics. Therefore, current biology poses new challenges pertaining to data management, querying and analyses.

As a result of high availability of intensive data stream information and owing to the advances made in computing technologies with high performance, big data analytics as on date emerged as high aspiring platform for real time descriptive and predictive analyses on large amounts of biological data, giving way for intelligent decisions making biology, a predictive science [6].

## 4 Proposed Framework

It is impossible without the competent computational resources with storage management and analysis of genomic information even if massive amounts of genomic data generated by high performance technologies [7]. Big data technologies enables storage and processing of very large data sets in distributed and parallel computing environment. Hence, It is significant to model the biological data to analyse and visualise the biological connections existed among them using big data modeling techniques.

Biological sequences are available in many formats like fasta, fastaq etc. Before performing analytics on biological sequences, we should convert them into common

format. By using Bio package in python one can extract the data from those formats and write them into CSV. Following code snippet shows how to extract the data from fasta file using Bio package.

```
from Bio import SeqIO  
with open("MySeq.fasta", "rU") as handle:  
for seq in SeqIO.parse(handle, "fasta"):  
print(seq.id)
```

After downloading the biological datasets there may be unwanted fields along with the sequences. After getting the data into CSV file, prioritizing the fields required for the application has to be done. After arranging the data into CSV file, biological sequences in big data environment can be analyzed.

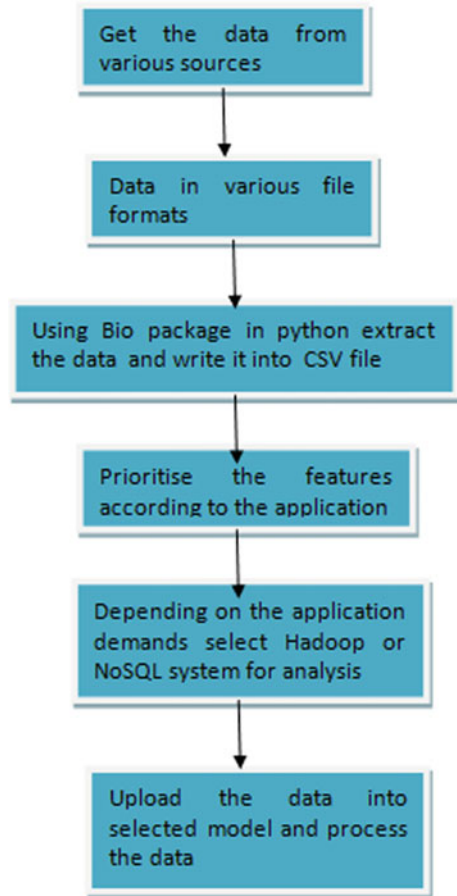
Hadoop MapReduce provides a powerful programming model to analyse the data. At the same time NoSQL databases having analytical power store huge amounts of data. Most of Document Based NoSQL databases support MapReduce programming model. MongoDB is compatible with MapReduce framework for aggregation. [8]. Besides, it is compatible with capped collection and execution of server side java script. Writing MapReduce code in Documents based databases is simple and flexible than Hadoop MapReduce. But only less computation can be done in NoSQL with MapReduce. However, NoSQL databases have real-time data extraction and processing at low latency. Hence one can select either Hadoop or NoSQL for analysis based on their application demands. Standard algorithms of biological sequence can be developed in MapReduce style like Needleman-Wunsch algorithm and algorithms for finding longest subsequence etc. [9, 10]. Figure 2 describes the proposed framework for biological sequence modeling and analysis.

Due to scalability nature of NoSQL databases or Hadoop Distributed File System we can produce the effective results for large scale problems.

## 5 Conclusion and Future Work

Not only is the big data analytics important in bioinformatics but also modeling big biological data. Because Data modeling provides a visual way to manage data sources creating data architecture so important that it increases the output and decrease the computing cost. In future we will extend this work through proper experimentation and results.

**Fig. 2** Framework for biological sequence modeling and analysis



## References

1. S.J. Bolla, S. Jyothi, Big data modeling for predicting side-effects of anticancer drugs: a comprehensive approach, in *Advances in Intelligent Systems and Computing*, vol. 1037 ed. by Y. Bi, R. Bhatia, S. Kapoor, (Springer, Cham, 2019)
2. C.J. Tauro, B.R. Patil, K.R. Prashanth, A comparative analysis of different NoSQL databases on data model, query model and replication model, in *Proceedings of International Conference on Emerging Research in Computing, Information, communication and Applications*, ERCICA (2013). ISBN: 9789351071020
3. I. Khan, S.K. Naqvi, M. Alam, S.A. Rizvi, Data model for big data in cloud environment, in *2015 IEEE* (2015). ISBN: 978-9-3805-4416- 8/15/\$31.00\_c
4. J.M. Ostell, S.J. Wheelan, J.A. Kans, *The NCBI data model in bioinformatics: a practical guide to the analysis of genes and proteins*, 2nd edn (Wiley, Hoboken, 2001), pp. 19–44. ISBN: 0471383910
5. Y. Li, L. Chen, Big biological data: challenges and Opportunities. *Genomics Proteomics Bioinform.* **12**, 187–189 (2014)



6. F. Ji, R. Elmasri, Y. Zhang, B. Ritesh, Z. Raja, Incorporating concepts for bioinformatics data modeling into EER models, in *3rd ACS/IEEE International Conference on Computer Systems and Applications*, vol. 2005, pp. 189–192 (2005). ISSN: 1387039 <https://doi.org/10.1109/aiccsa.2005.1387039>
7. G.M. Siddesh, K.G. Srinivasa, I. Mishra, A. Anurag, E. Uppal, Phylogenetic analysis using mapreduce programming model, in *IEEE International Parallel and Distributed Processing Symposium Workshops* (2015)
8. DataFlair team. <https://data-flair.training/blogs/hadoop-vs-mongodb/>
9. K.E. Kannammal, C.P. Shabariram, Dna global sequence alignment using map reduce in openstack sahara. Kannammal et al. *Int. J. Adv. Eng. Technol*
10. N.P. Kandel, S.R. Joshi, A map-reduce model to find longest common subsequence using non-alignment based approach, in *Proceedings of IOE Graduate Conference*, pp. 329–336 (2016)

# Specification and Estimation of a Biometric Model by Using Logistic Regression for Measuring Child Mortality



P. Vishnu Priya, B. Sarojamma, G. Madhusudan, P. Srivyshnavi, M. Naresh, and P. Balasiddamuni

**Abstract** Children are the most valuable assets of the Nation. The social and economic development of the Nation has been based on welfare of the children in the Nation. The level of child mortality in a country is not only an indicative of the public health but, it also can be considered as an index of quality of life lived by the people in the country. The various factors influencing child mortality rate are given by: ecological factors, health status variables, vital characteristics, socio-economic characteristics, cultural factors and others. Simple Logistic Regression models can be specified and the estimates of their parameters can be used to measure child mortality in Biostatistics. Generally, Maximum likelihood method of estimation, Non-weighted least squares estimation for the analysis of Bio-categorical data and Discriminant function analysis can be used to estimate the parameters of logistic regression model. In this paper, an attempt has been made by specifying and estimating biometrical models based on Multivariable Logistic Regression Model and Multinomial Logistic Regression Model.

**Keywords** Mortality rate of child · Logistic regression model · Maximum likelihood estimation

## 1 Introduction

Statistical Science provides various statistical tools which can have their applications in different fields of Science. Some important branches of Statistical Science exist in the literature are Biometrics, Demometrics, Technometrics, Psychometrics, Econometrics Operations Research, Criminometrics, Anthropometrics, Agricultural Statistics, Economic Statistics, Medico Statistics, Business Statistics, Industrial Statistics, Mathematical Statistics, Statistical Physics and Psephology etc. Biometrics is a

---

P. Vishnu Priya (✉) · B. Sarojamma · G. Madhusudan · M. Naresh · P. Balasiddamuni  
Department of Statistics, Sri Venkateswara. University, Tirupati 517502, India  
e-mail: [vishnu.pasala@gmail.com](mailto:vishnu.pasala@gmail.com)

P. Srivyshnavi  
Department of CSE, S.P.M.V.V. Engineering College, Tirupati 517502, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_37](https://doi.org/10.1007/978-3-030-46939-9_37)

rapidly expanding field of Statistical science. Biometrical methods have a wide number of applications in the fields of Biology, Demography, Epidemiology, Anthropology, Medicine and many others. Quantitative genetics including population genetics and Demometrics can be considered as one of the main branches of Biometrics. Mortality is the main component of Demometrics that studies rates and causes of deaths for a population as a whole. Infant Mortality Rate (IMR) and Child Mortality Rate (CMR) have special significance in Health statistics because, they are considered to be most sensitive indices of health conditions of the general population. The computations of IMR and CMR on the basis of the infants population and children population respectively present difficulty because, the infants and children are generally under enumerated. Very often, the estimates of the population by age may not be available. Child Mortality has been caused by several factors (known as Life-Affecting variables) which have been listed out in a few conceptual models. The main contributions in estimating conceptual Models for Child and Infant Mortality Rates were made by Sullivan [1], Ruzika and Kanitkar [2], Trussell [3], Feeney [4], Meegama [5], Mosley and Chen [6], Mahadevan et al. [7, 8], Chen [9], Jain and Pravin [10], Quamrul et al. [11], Pedersen [12] and others. Children are the most valuable assets of the Nation. The social and economic development of the Nation has been based on welfare of the children in the Nation. A welfare state aims to include at least the right of every child to be given an equal chance to live, to be educated and to be enabled to develop its skills commensurate with its ability to attain adulthood and the right to work, and to have a minimum standard of life. The level of child mortality in a country is not only an indicative of the public health but, it also can be considered as an index of quality of life lived by the people in the country.

## 2 Child Mortality and Related Influencing Factors

The various factors influencing child mortality rate are given by:

**Cultural Factors:** Customs ( $X_{11}$ ) Beliefs and values ( $X_{12}$ ), Women status ( $X_{13}$ ), Women role in family ( $X_{14}$ ) etc.

**Ecological Factors:** House conditions ( $X_{21}$ ), House Sanitation ( $X_{22}$ ) Water supply ( $X_{23}$ ), Type of house ( $X_{24}$ ) etc.

**Socio-Economic Characteristics:** Education ( $X_{31}$ ), Occupation of the parents ( $X_{32}$ ), Family Income ( $X_{33}$ ), Socio-economic Status ( $X_{34}$ ).

**Demographic Characteristics:** Marriage age ( $X_{41}$ ), Mother's age at 1st birth ( $X_{42}$ ), Mother's number of live births ( $X_{43}$ ), Mother's Closed Birth Interval ( $X_{44}$ ), Spacing of births ( $X_{45}$ ), Age at child deaths ( $X_{46}$ ) etc.

**Health Status Variables:** Health status of women ( $X_{51}$ ) Type of medicine taken during pregnancy ( $X_{52}$ ), Medical check-up during pregnancy ( $X_{53}$ ), Type of treatment given during pregnancy ( $X_{54}$ ), Type of delivery attendants ( $X_{55}$ ) etc.

Biometrical models for estimating Child Mortality Rate identified the aforementioned variables as independent variables, which influence the child Mortality Rate.

### 3 Child Mortality Rate

**Child Mortality:** It is the chance of dying between the 1st and 5th birthday (deaths 1 to <5 years of old children).

**Child Mortality Rate (CMR):** It is defined as the ratio of child deaths (between 1 and <5 years) to the live births (from 1 to <5 years) multiplied by 1000.

$$CMR = \left[ \frac{\text{Child deaths between 1 and 5 years}(D_{1\text{ to } <5})}{\text{Live births from 1 to under 5 years}(B_{1\text{ to } <5})} \right] 1000$$

**Under 5 Mortality (U5M):** The chance of dying the 5th birthday.

Generally, all the mortality rates are expressed before as deaths per 1000 live births, except Child Mortality Rate, which is expressed as deaths per 1000 children surviving to one year.

### 4 Specification of Multivariable Logistic Regression Model for Child Mortality

Consider a Logistic Regression relationship between child mortality and the following set of influencing variables:

- $x_1$ : Cultural factors variable
- $x_2$ : Ecological factors variable
- $x_3$ : Socio-economic factors variable
- $x_4$ : Demographic factors variable and
- $x_5$ : Health status factors variable.

The Multivariable Logistic Regression Model for Child Mortality [ $P(x)$ : Chance of dying between exact ages 1 and 5 years] and ‘m’ influencing variables ( $m = 5$ ) can be specified as

$$P(x) = \frac{e^{G(x)}}{1 + e^{G(x)}}$$

where  $x$  is a vector of  $m$  influencing variable. The Logit of the Multivariable Logistic Regression model is given by:

$$G(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

Here, some of the influencing variables may be discrete, nominal scale variables and other variables may be internal scale variables. The Multivariable Logistic Regression model can be estimated by using the Method of Maximum Likelihood estimation. For a sample of ‘n’ independent observations  $(X_{ij}, Y_i), i = 1, 2, \dots, n; j = 1, 2, \dots, m$ , on m influencing variables and child mortality variable, one may fit this model for child mortality as in the case of univariate logistic regression model for child mortality and find the estimators of parameters of the model. The maximum likelihood equations can be written as,

$$\sum_{i=1}^n [Y_i - P(X_i)] = 0 \text{ and } \sum_{i=1}^n X_{ij}[Y_i - P(X_i)] = 0, \quad j = 1, 2, \dots, m.$$

The maximum likelihood estimator of parametric vector  $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)$  may be denoted by  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m)$  and it is obtained by solving the ML equations. The information matrix of  $\hat{\beta}$  is given by  $\hat{I}(\hat{\beta}) = X^T \hat{V} X$  Where  $X$  is  $n \times (m + 1)$  matrix of influencing variables.  $\hat{V}$  is an  $(n \times n)$  diagonal matrix with general element  $[\hat{P}(x_i)(1 - \hat{P}(x_i))]$  i.e., The matrices  $X$  and  $V$  are given by

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1m} \\ 1 & X_{21} & X_{22} & \dots & X_{2m} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix}_{n \times (m+1)}$$

$$\hat{V} = \begin{bmatrix} \hat{P}_1(1 - \hat{P}_1) & 0 & \dots & 0 \\ 0 & \hat{P}_2(1 - \hat{P}_2) & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & \hat{P}_n(1 - \hat{P}_n) \end{bmatrix}_{n \times n}$$

Here,  $\hat{P}_i = \hat{P}(X_i), i = 1, 2, \dots, n$ .

The estimated variance covariance matrix of estimates of regression coefficients can be obtained as  $\widehat{\text{Var}}(\hat{\beta}) = [\hat{I}(\hat{\beta})]^{-1}$ . The estimated SE of the estimates of regression parameters are given by  $\widehat{\text{SE}}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}, j = 0, 1, 2, \dots, m$ . To test for the  $H_o = \beta_j = 0$ , wald test statistic is given by  $W_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim N(0, 1)$ .

The Multivariate analog of Wald test statistic for the multivariable logistic regression model is given by,  $W = \hat{\beta}' [\widehat{\text{Var}}(\hat{\beta})]^{-1} \hat{\beta}$  or  $W = \{\hat{\beta}' [X' V X] \hat{\beta}\} \sim \chi^2_{(n+1)}$ .

## 5 The Multinomial Logistic Regression Model for Child Mortality

Suppose that child mortality is nominal with more than two levels. Consider the influencing variables or independent variables or covariates as

- (i) Cultural factors variable ( $X_1$ )
- (ii) Ecological factors variable ( $X_2$ )
- (iii) Socio-economic factors variable ( $X_3$ )
- (iv) Demographic factors variable ( $X_4$ ) and
- (v) Health status factors variable ( $X_5$ ).

One may assume, Discrete dependent variable  $Y$  are coded  $0, 1, 2, \dots, k$ ; and a collection of  $(m + 1)$  influencing variables as the vector  $X^1 = (X_0, X_1, X_2, \dots, X_m)$  where  $X_0 = 1$ . In the case of  $(k + 1)$  category discrete dependent variable logistic regression model, one may need ' $K$ ' 'Logit Functions'. Generally, one may use  $Y = 0$  as the reference dependent variable and obtain Logits comparing  $Y = 1, Y = 2, \dots, Y = k$ .

The  $K$  Logit functions are given by

$$g_j(X) = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \dots + \beta_{jm}X_m \text{ or } g_j(X) = X' \beta_j, \quad j = 1, 2, \dots, k$$

Now, the Multinomial or Discrete choice or polychotomous or polytomous Logistic Regression model can be specified as

$$P_i(X) = P(Y = i/x) = \frac{\exp(g_j(X))}{\sum_{q=0}^K \exp(g_q(X))}, \quad i = 0, 1, 2, \dots, k$$

$$\Rightarrow P(i/x) = \frac{1}{1 + e^{g_1(X)} + e^{g_2(X)} + \dots + e^{g_k(X)}}, \quad i = 0, 1, 2, \dots, k$$

Here, conditional probability of each dependent variable category given the influencing variables vector is a function of the vector of  $K(m + 1)$  coefficients namely,

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}$$

Here, each  $\beta_j = 1, 2, \dots, k$  is a vector of  $(m + 1)$  parameters. It should be noted that the vector  $\beta_0 = 0$  and  $g_0(x) = 0$ . The parametric vector  $\beta$  is estimated by using MLE method. The likelihood function may be written as

$$\phi(\beta) = \prod_{s=1}^n P_0(x_s)^{Y_{0s}} P_1(x_s)^{Y_{1s}} P_2(x_s)^{Y_{2s}}, \dots, P_k(x_s)^{Y_{ks}}$$

Taking the logarithm and using  $\sum Y_{js} = 1$  for each  $s$ , Logarithmic likelihood function can be written as

$$\begin{aligned} \text{Ln}\phi(\beta) = L(\beta) &= \sum_{s=1}^n [Y_{1s}g_1(x_s) + Y_{2s}g_2(x_s) + \dots + Y_{ks}g_k(x_s)] \\ &\quad - \text{Ln}[1 + e^{g_1(x_s)} + e^{g_2(x_s)} + \dots + e^{g_k(x_s)}] \end{aligned}$$

The maximum likelihood equations can be obtained by taking the first order partial derivatives of  $L(\beta)$  w.r.t.  $K(m + 1)$  coefficients.

By denoting  $P_{js} = P_j(X_s)$ , the general MLE can be expressed as

$$\frac{\partial L(\beta)}{\partial \beta_{jr}} = 0 \Rightarrow \sum_{s=1}^n X_{rs}(Y_{js} - P_{js}) = 0$$

For  $j = 1, 2, \dots, k$  and  $r = 0, 1, 2, \dots, m$ , with  $X_{0s} = 1$ , for all  $s = 1, 2, \dots, n$

The ML estimates of  $\beta$  can be obtained by solving the ML equations, using an Iterative Numerical analysis technique. To obtain the Information matrix  $I(\hat{\beta})$ . The elements in matrix of 2nd order partial derivatives are given by

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jr} \partial \beta_{j^1 r^1}} = - \sum_{s=1}^n X_{r^1 s} X_{rs} P_{js} (1 - P_{js}) \text{ and } \frac{\partial^2 L(\beta)}{\partial \beta_{jr} \partial \beta_{j^1 r^1}} = \sum_{s=1}^n X_{r^1 s} X_{rs} P_{js} P_{j^1 s}$$

$\forall j$  and  $j^1 = 1, 2, \dots, k$  and  $\forall r$  and  $r^1 = 0, 1, 2, \dots, m$ .

The observed information matrix  $I(\hat{\beta})$  is the  $K(m + 1)$  by  $K(m + 1)$  matrix, whose elements are the negatives of the values in equations and evaluated at  $\hat{\beta}$ . The estimated covariance matrix of the maximum likelihood estimator  $\hat{\beta}$  is given by

$$\widehat{\text{Var}}(\hat{\beta}) = [I(\hat{\beta})]^{-1}$$

Suppose that the matrix  $X$  be  $[n \times (m + 1)]$  matrix containing the values of the Influencing variables; and matrix  $V_j$  be the  $(n \times n)$  diagonal matrix with general element  $\hat{P}_{ji}(1 - \hat{P}_{ji})$  for  $j = 1, 2, \dots, k$ .

The estimator of the Information matrix is obtained as

$$\hat{I}(\hat{\beta}) = \begin{bmatrix} \hat{I}(\hat{\beta})_{11} & \hat{I}(\hat{\beta})_{12} & \dots & \hat{I}(\hat{\beta})_{1k} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \hat{I}(\hat{\beta})_{k1} & \hat{I}(\hat{\beta})_{k2} & \dots & \hat{I}(\hat{\beta})_{kk} \end{bmatrix}$$

where  $\hat{I}(\hat{\beta})_{11} = (X^1 V_1 X)$ ,  $\hat{I}(\hat{\beta})_{22} = (X^1 V_2 X)$ ,  $\dots$ ,  $\hat{I}(\hat{\beta})_{kk} = (X^1 V_k X)$   
 and  $\hat{I}(\hat{\beta})_{jj^1} = \hat{I}(\hat{\beta})_{j^1 j} = (X^1 V_{jj^1} X)$ .

where  $V_{jj^1}$  is the diagonal matrix with element  $\hat{P}_{js} \hat{P}_{j^1 s}$ ,  $s = 1, 2, \dots, n$ .

For the significance of estimated regression coefficients, the odds ratios can be estimated as follows: The Odds Ratio of dependent variable  $Y = j$  versus  $Y = 0$  for Influencing variable values of  $x = a$  versus  $x = b$  is given by

$$O_{R_j}(a, b) = \frac{P(Y = t/x = a)/P(Y = o/x = a)}{P(Y = t/x = b)/P(Y = o/x = b)}, \quad \forall t = 1, 2, \dots, k.$$

It should be noted that the estimating and interpreting Child Mortality from proposed model is almost equal in Dichotomous situation and difference may be observed in defining Odds Ratio for Influencing Variables in Multinomial case.

## 6 Conclusions

The Biometrical analysis of demographic data is of vital interest to policy planners and health administrators in the preparation and implementation of development strategies to meet the health needs and demands of people. It is essential in implementation and evaluation of public health programmes. Many experts in the Demometrics field have been argued that child Mortality Rate is an significant indicator of the socio-economic condition, health and nutritional standard of a community. Child Mortality estimation from proportion of dead among children ever born to women in standard age groups is now a familiar technique. Several demographic studies have shown that child mortality influenced by a number of socio economic variables. In



the present study, the various variables influencing child mortality have been classified into five groups namely, (i) cultural variables (ii) Ecological or Environmental variables (iii) Socio-economic variables (iv) Demographic variables and (v) Health status variables. A multivariable logistic regression model for child mortality and its inferential aspects have been proposed to estimate child mortality in the present research work. Further, multinomial logistic Regression model for child mortality involving five important influential variables has been developed in the present study.

## References

1. J. Sullivan, Models for the estimation of the probability of dying between birth and exact ages of early childhood. *Popul. Stud.* **26**, 79–98 (1972)
2. L.T. Ruzika, T. Kanitkar, Infant mortality in greater Bombay. *Demography India* **2**(1), 41–55 (1973)
3. I.J. Trussell, A re-estimation of the multiplying factors for the brass technique for determining childhood survival rates. *Popul. Stud.* **29**(1), 97–108 (1975)
4. G. Feeney, Estimating infant mortality trends from child survivorship data. *Popul. Stud.* **34**(1), 109–128 (1980)
5. S.A. Meegama, *Socio-Economic Determinants of Infant and Child Mortality in Sri Lanka: An Analysis of Post War Experience* (International Statistics Institute, Voorburg, Netherlands, 1980)
6. W.H.A. Mosley, C.L. Chen, *Child Survival: Strategies for Research* (Cambridge University Press, New York, 1983)
7. K. Mahadevan et al., *Infant and Child Mortality in India* (Mittal Publishers, New Delhi, 1986a)
8. K. Mahadevan et al., *Infant Mortality and Life Affecting Variables* (Mimeo, New York, 1986b)
9. C.L. Chen, Child survival: levels, trends and determinants in determinants of fertility in developing countries; a summary of knowledge, Part A, in *Committee on Population and Demography* (National Academy Press, Washington, DC, 1983), pp. 163–190
10. A.K. Jain, V. Pravin, *Infant Mortality in India* (Sage Publications, London, 1987)
11. H.C. Quamrul, M.D. Rafiqul Islam, K. Hossain, Effects of demographic characteristics on neonatal, post neonatal, infant and child mortality. *Curr. Res. J. Biol. Sci.* **2**(2), 132–138 (2010)
12. J. Pedersen, J. Liu, Child mortality estimation: full birth histories. *PLOS Med. J.* 1001289 (2012)
13. R. Garfield, C.-S. Leu, A multivariate method for estimating mortality rates among children under 5 years from health and social indicators in Iraq. *Int. J. Epidemiol.* **29**, 510–515 (2000)
14. K. Hill, Approaches to the measurement of childhood mortality: a comparative review. *Popul. Index* **57**(3), 368–382 (1991)
15. D.W. Hosmer, S. Lemeshow, *Applied Logistic Regression*, 2nd edn. (Wiley, New York, 2000)
16. M.C. Paik, The generalized estimating equation approach when data are not missing completely at random. *J. Am. Stat. Assoc.* **92**, 1320–29 (1997)
17. A. Palloni, A new technique to estimate infant mortality with an application for El Salvador and Colombia. *Demography* **16**(3) (1979)
18. M.M. Rahman, M.R. Islam, M.K. Ali, Effects of demographic characteristics on infant and child mortality: a case study of Rajshahi district. *Bangladesh* **12**(2), 161–173 (2005)
19. C.R. Rao, *Linear Statistical Inference and Its Applications*, 2nd edn. (Wiley, New York, 1973)
20. R. Silva, Child mortality estimation: consistency of under-five mortality rate estimates using full birth histories and summary birth histories. *PLOS Med. J.* 1001296 (2012)
21. P. Vishnu Priya, Statistical techniques for estimating the child mortality rate. Unpublished Ph.D. thesis in Statistics, S.V. University, Tirupati, (2018)

# A Case Study Report: Ruptured Scar Ectopic Pregnancy



Abhilaasha Macherla and R. V. Raviteja

**Abstract** Caesarean Scar Pregnancy (CSP) is a rare type of ectopic pregnancy related to increasing incidence of caesarean sections that are diagnosed and reported. Here we present report of case study of ruptured caesarean scar site in ectopic pregnancy admitted in the emergency department in relation with hypovolemic shock. Requirement for emergency laparotomy was based on the clinical background and on the report of ultrasonographical studies. Severe life threatening conditions are related with such CSP case due to rupture in the uterine wall leading to haemorrhage and maternal mortality.

**Keywords** Caesarean scar pregnancy (CSP) · Hysterectomy · Ultrasonography · Ectopic pregnancy

## 1 Introduction

Caesarean Scar ectopic pregnancy is implantation of fertilized egg in the myometrium of the scar formed at the previous caesarean section. It is a rare case of ectopic pregnancy. The incidence is 1 in 2226 of all pregnancies and 0.15% previous caesarean cases. Untreated case of Caesarean Scar Pregnancy is life threatening due to rupture and haemorrhage of the uterine wall with increased chances of maternal mortality. As prevention is better than cure treating the Uterine scar as early as diagnosed is far better than removal of Caesarean Scar Pregnancy by Laparotomy and Hysterectomy.

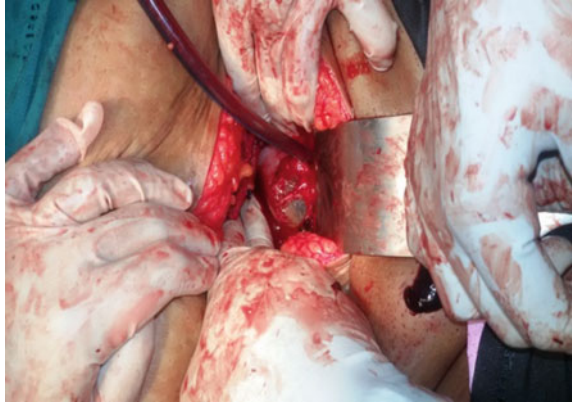
---

A. Macherla (✉) · R. V. Raviteja  
Genecologist, M.Ch. (Urology), Guntur Medical College, Guntur, India  
e-mail: [abhilaashart@gmail.com](mailto:abhilaashart@gmail.com)

R. V. Raviteja  
e-mail: [raviteja091@gmail.com](mailto:raviteja091@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_38](https://doi.org/10.1007/978-3-030-46939-9_38)

**Fig. 1** Visualization of amniotic sac with live fetus through ruptured scar



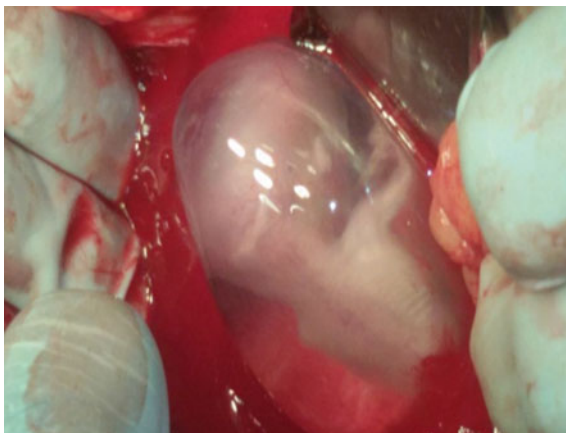
## 2 The Case Study

This is a case study of a unregistered Pregnant woman aged 30 with G2P1L1 came in emergency with abdominal pain, vomiting from past 3 days followed by vaginal bleeding since an hour. Her last menstrual period was on 26-7-2015 with previous regular cycles and gestational period of 10 Weeks 3 days. Her Urine pregnancy test was positive one month ago with a previous history of Full term caesarean section 5 years ago. During her gestational period, she had not undergone any antenatal check-up. On examination she had tachycardia, Hypotension with moderate pallor. On further examination via Speculum and Bimanual method it was found that uterus seemed enlarged with slight bleeding through Cervical Os. Her TVS scan showed Single live intrauterine gestation of 10 weeks 5 days with mixed echoic mass lesion in the pelvis anterosuperior to the uterus predominantly on left side with moderate ascites showing internal echoes—suggestives of ruptured left tubal ectopic gestation. Ultrasound guided aspiration of ascites revealed haemorrhagic fluid. Her haemoglobin was 7.5 gm. Patient was resuscitated with IV fluids. IV antibiotics started. Patient shifted for exploratory laparotomy after taking consent. Intra-operatively 2 L haemoperitoneum was found and there was ruptured uterine caesarean scar where Amniotic Sac along with live fetus was protruding out (Figs. 1 and 2). Intact Sac removed (Fig. 3). Uterine contents emptied and defect repaired in two layers (Fig. 4). 3 units of packed cells and 3 units FFP were transfused intra-operatively and post-operatively.

## 3 Discussion

This is a case report of ruptured scar ectopic where the TVS misdiagnosed as left ruptured tubal pregnancy. CSP is implantation of fertilized egg in the uterine myometrium layer of the scar formed in the previous caesarean section, a rare case of ectopic pregnancy. The incidence is 1: 2226 of all pregnancies with 0.15% women

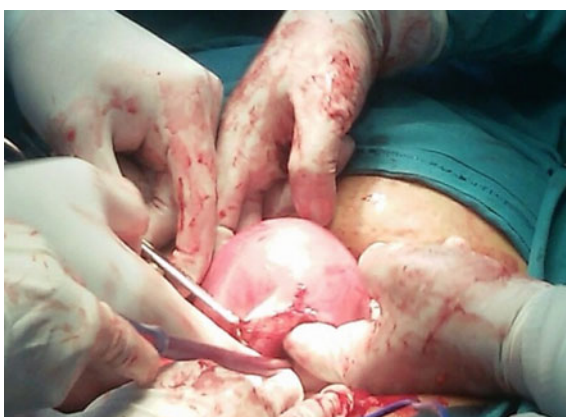
**Fig. 2** Intact sac with live fetus



**Fig. 3** Intact sac removal



**Fig. 4** Evacuation of uterus



with previous caesarean. And among all ectopic pregnancies 6.1% pregnancies are found to be scar ectopic in women with at least one caesarean delivery. Ultrasonography stand as a primary line of diagnostic tool for such type of scar pregnancy with sensitivity of 86.4%. The Criteria for the diagnosis are Empty uterus and cervical canal with clearly visible endometrium, Presence of gestational sac within the anterior portion of the lower uterine segment at the presumed site of the scar, thinning or absent of myometrium between the gestational sac and bladder. The Conservative management is with systemic or local Methotrexate inj., dilatation and curettage of gestational sac, Hysteroscopy, Laparoscopy, surgical laparotomy/hysterectomy, and uterine artery embolization [1, 2]. A delay in diagnosis may lead to uncontrolled haemorrhage of uterus which may require hysterectomy [3, 4].

## 4 Conclusion

Caesarean Scar Pregnancy (CSP) is rare complication in the pregnancy that is best diagnosed by transvaginal ultrasonography. A delay in diagnosis may lead rupture of uterine layers that may further require Hysterectomy. However, early diagnosis using ultrasonography can lead to conservative measures that prevents maternal mortality thus preserving subsequent fertility.

## References

1. A. Ash, A. Smith, D. Maxwell, Caesarean scar pregnancy. *Br. J. Obstet. Gynaecol.* **114**(3), 253–263 (2007)
2. D.L. Fylstra, T. Pound-Chang, M.G. Miller, A. Cooper, K.M. Miller, Ectopic pregnancy within a cesarean delivery scar: a case report. *Am. J. Obstet. Gynecol.* **187**(2), 302–304 (2002)
3. M.A. Rotas, S. Haberman, M. Levгур, Cesarean scar ectopic pregnancies. *Obstet. Gynecol.* **107**(6), 1373–1381 (2006)
4. R. Maymon, R. Halperin, S. Mendlovic, D. Schneider, A. Herman, Ectopic pregnancies in a caesarean scar: review of the medical approach to an iatrogenic complication. *Hum. Reprod. Update* **10**(6), 515–523 (2004)

# Some Modified Biometrical Diversity and Evenness Indices



G. Madhusudan, P. Srivyshnavi, B. Sarojamma, M. Naresh, R. Abbaiah, and P. Balasiddamuni

**Abstract** Biometrical Genetics is one of the vital branches of genetic science which has become particularly essential for the Biological, Agricultural and Medical Scientists in their research aspects. Biological Diversity and Evenness are very vital concepts in Biometrical Genetics. A large number of measures such as Diversity and Evenness indices represented by a species abundance distribution have been proposed in recent literature on ecological research. Diversity in the species can be measured in many ways. Biometricians recognize three components of diversity to describe and compare different communities given by Alpha, Beta and Gamma diversities. In the present recent article, some modified biological diversity and evenness indices have been proposed to study the species abundance distribution.

**Keywords** Diversity · Evenness · Types of diversities

## 1 Introduction

Biological Diversity is a fertile new research field of science and its patterns are created by ecological, evolutionary and conservation processes. Conceptually, the word 'Biodiversity' was introduced in the Mid-1980s and it captures the essence of research relating to total richness and variety of life on earth. There are mainly four types of Biodiversity namely, Species Diversity, Genetic Diversity, Ecosystem Diversity and Functional Diversity. The two main components of Biodiversity are richness and evenness. Species richness is expressed as the number of species and evenness is expressed as the proportion of species or functional groups present on a site. The greater evenness of a site relates the more equal species are in proportion to each other, and the low evenness site indicates that a few species dominate the site.

---

G. Madhusudan (✉) · B. Sarojamma · M. Naresh · R. Abbaiah · P. Balasiddamuni  
Department of Statistics, S.V. University, Tirupati, India  
e-mail: [gmadhusudan7@gmail.com](mailto:gmadhusudan7@gmail.com)

P. Srivyshnavi  
Department of CSE, S.P.M.V.V. Engineering College, Tirupati, India

Biodiversity has importance in human life and medicine. It offers food harvests, fish, livestock etc. for human and a large number of species of plants has been used for medicine purposes since, very ancient times.

## 2 Measurement of Biodiversity

Biometrical diversity is a vital concept in quantitative ecology that has been extensively studied by biostatisticians for over six decades. A mathematical measure of species diversity in a given community is known as Biodiversity Index. Quantifying Diversity appears as simple and unambiguous concept than quantifying Evenness. A large number of measures such as Diversity and Evenness indices represented by a species abundance distribution have been proposed and well established in recent literature on Ecological research. Diversity in the species can be measured in many ways. Biostatisticians recognize three components of Diversity to describe and compare different communities which are given by (i) Alpha (ii) Beta and (iii) Gamma Diversity. Alpha diversity is local diversity; Gamma diversity is the total regional diversity of large area and Beta Diversity links Alpha and Gamma diversities or local and regional diversities. The general measures of central tendency or measures of dispersion are not applicable to study the dispersion or variation with reference to the Nominal scale data. One may involve the concept of Diversity, the distribution of indistinguishable observations among categories similar to the concept of Dispersion. If the observations distributed evenly among categories, then it shows high Diversity. If a set of observations, where the bulk of the data; occurs in very few of the categories, then it exhibits low diversity.

### 2.1 Biodiversity Measures

Most of the Biological Diversity measures are developed by using Information Theory underlying the concept of uncertainty or probability. Some important Biodiversity indices, which are available in the literature are given by:

- (i) Shannon's [1] entropy measure of Biological Diversity ( $H_S$ ): Bowman et al. [14] proved that,  $H_S$  is an under estimate of the population Diversity measure. However the bias in the estimation decreases with increasing the sample size.
- (ii) Simpson's [2] Biological Diversity Index ( $H_{Sim}$ )
- (iii) Brillouin's [3] Biological Diversity Index ( $H_B$ ): It is an Information Theoretic Diversity measure.
- (iv) Pielou's [4] Biological Diversity Index ( $H_P$ ): It is based on  $H_S$ , which is the ratio or proportion of  $H_S$  to Its maximum possible diversity  $H_S(\text{Max})$ . Sometimes,  $H_P$  is known as Evenness Index or Homogeneity Index or Relative

Diversity Index. The measure  $H_p^1 = [1 - H_p]$  may be considered as Heterogeneity index or Dominance index. It should be noted that  $H_p$  is an over estimate of the Population Evenness measure.

- (v) Sheldon's [5] Biological Diversity Index ( $H_{Shel}$ ).
- (vi) Renyi's [6] Generalized Entropy measure of Biological Diversity ( $H_R$ ).
- (vii) Hills' [7] Family of Biological Diversity Indices ( $H_{H(\alpha)}$ ): It is antilog of Renyi's Generalized Entropy measure of Biological Diversity.
- (viii) Heip and Engel's [8] Biological Diversity Index or Corrected Sheldon's Biological Diversity Index ( $H_H$ ).
- (ix) Patil and Taillie [9, 10] Biological Diversity Index ( $H_\beta$ ).
- (x) T Sallis [11] Non-Expensive Entropy measure of Biological Diversity ( $H_{T.S(\alpha)}$ ).

### 3 Biological Evenness Measures

A large number of Biological Evenness (Equitability) indices available in the literature were reviewed by Smith and Wilson [12]. Among them, they preferred four important Evenness indices, which have more practical importance, are given by:

- (i) Simpson's Biological Evenness Index ( $E_{Sim}$ ): It should be noted that Simpson's Evenness Index ranges from 0 to 1 and is relatively unaffected by the rare species in the sample.
- (ii) Camargo's [13] Biological Evenness Index ( $E_C$ ): This index is unaffected by categories or species richness. It is very simple to calculate, compared with other Evenness indices. Generally,  $E_C$  is relatively little affected by the rare species in the sample.
- (iii) Smith and Wilson's [12] Biological Evenness Index ( $E_{SW}$ ): This index is independent of species richness and it is sensitive to both rare and common to species in the community.
- (iv) Modified Nee et al. Biological Evenness Index ( $E_{MN}$ ): This index ranges from 0 to 1 and is independent of species richness. It is sensitive to both common and rare species in the sample.

### 4 Modified Biometrical Diversity and Evenness Indices

There are mainly two types of Diversity Indices namely Type-I and Type-II Indices. Type-I Diversity Indices are most sensitive to changes in the rare species or categories in the community sample. Shannon's Diversity Index is an example of Type-I Diversity Index. Type-II Diversity Indices are most sensitive to changes in the more abundant species (or categories). Simpson's Diversity Index in an example of Type-II



Diversity Index. Evenness Indices can be obtained by scaling the Diversity measure or heterogeneity measure relative to its maximum value then each category or species in the sample is represented by the same number of individuals. Evenness Indices can be obtained by scaling the Diversity measure or Heterogeneity measure relative to its maximum value when each category or species in the sample is represented by the same number of individuals. Its range or arithmetic mean is of maximum and minimum value.

**Notations:** Consider the species (category) abundance distribution of the community (Population) as a vector  $\underline{N} = (N_1, N_2, \dots, N_K)$  such that  $N = \sum_{i=1}^K N_i$ . The species (categories) proportions of the community (Population) are given by a vector  $\underline{P} = (P_1, P_2, \dots, P_k) = (N_1/N, N_2/N, \dots, N_k/N)$ , Generally,  $N_i$  and  $P_i$  are unknown,  $\forall i = 1, 2, \dots, k$ . Here  $N_i$  is the number of individuals of species (categories)  $i$  in the region (population) interest. The species (categories) sample proportions are given by a vector  $\hat{\underline{p}} = (p_1, p_2, \dots, p_k) = (\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_k}{n})$ .

Such that  $n = \sum_{i=1}^K n_i =$  Total sample size.

For a perfectly Even Community (Population), one may have  $\underline{P}^* = (P_1^*, P_2^*, \dots, P_K^*) = (1/k, 1/k, \dots, 1/k)$ . Consider the Brillouin Information-Theoretic Diversity measure as

$$H_B = \frac{\text{Log} \left[ \frac{n!}{n^k (n!)^k} \right]}{n} \text{ or } H_B = \frac{\text{Log} \left[ \frac{n!}{(n_1!)(n_2!) \dots (n_k!)} \right]}{n} \text{ or } H_B = \frac{\text{Log}(n!) - \sum_{i=1}^k (n_i!)}{n}$$

for large  $n$  (i.e.,  $n \rightarrow \infty$ ) consider Stirling's approximation for  $n!$  as  $n! \simeq \sqrt{2\pi n} n^{n+1/2} e^{-n}$  as  $n \rightarrow \infty$ .

More precisely,  $\sqrt{2\pi n} n^{n+1/2} e^{-n} < n! < \sqrt{2\pi n} n^{n+1/2} (1 + \frac{1}{4n}) e^{-n}$ .

Using Stirling's approximation, one may obtain

$$\log n! = (n + 0.5) \log n - 0.434294n + 0.399090$$

An approximation with only half the error is given by  $n! \simeq \sqrt{2\pi} \left(\frac{n+0.5}{e}\right)^{n+0.5}$ .

And one may have,  $\log n! = (n + 0.5) \log(n + 0.5) - 0.434294(n + 0.5) + 0.399090$ .

By substituting the value for  $\text{Log}(n!)$  the Diversity Index  $H_B$  written as

$$H_B^* = \frac{[(n + 0.5) \log n - 0.434294n + 0.399090] - \sum_{i=1}^k [(n_i + 0.5) \log n_i - 0.434294n_i + 0.399090]}{n}$$

$$H_B^* = \frac{[(n + 0.5) \log n - \sum_{i=1}^k (n_i + 0.5) \log n_i + 0.399090(1 - K)]}{n}$$

$$H_B^{**} = \frac{[(n + 0.5) \log(n + 0.5) - 0.434294(n + 0.5) + 0.399090] - \sum_{i=1}^k [(n_i + 0.5) \log(n_i + 0.5) - 0.434294(n_i + 0.5) + 0.399090]}{n}$$

$$H_B^{**} = \frac{[(n + 0.5) \log(n + 0.5) - \sum_{i=1}^k (n_i + 0.5) \log(n_i + 0.5) + 0.181943(1 - K)]}{n}$$

The maximum possible value Brillouin Diversity Index for a set of  $n$  observations distributed among  $k$  categories or species is given by

$$H_B^{(Max)} = \frac{\log(n!) - (k - d) \log(q!) - d \log[(q + 1)!]}{n}$$

where  $q$  is the integer portion of  $(n/k)$  and  $d$  is the remainder.

Using Sterling's approximation, one may obtain

$$H_B^{*Max} = \frac{[(n + 0.5) \log n - 0.434294n + 0.399090] - (k - d)[(q + 0.5) \log q - 0.434294(q + 1) + 0.3990]}{n}$$

$$H_B^{*Max} = \frac{[(n + 0.5) \log n - (k - d)(q + 0.5) \log q - d(q + 1.5)] \log(q + 1) + 0.434294(qk + d - n) + 0.399090(1 - k)}{n}$$

$$H_B^{**Max} = \frac{[(n + 0.5) \log(n + 0.5) - 0.434294(n + 0.5) + 0.399090] - (k - d)[(q + 0.5) \log(q + 0.5) - 0.434294(q + 0.5) + 0.399090] - d[(q + 1.5) \log(q + 1.5) - 0.434294(q + 1.5) + 0.399090]}{n}$$

The minimum possible values of  $H_B$ ,  $H_B^*$  and  $H_B^{**}$  are given by zero

$$(H_B^{Min} = H_B^{*(Min)} = H_B^{**(Min)} = 0)$$

Now, the modified Brillouin Evenness Indices based on  $H_B^*$ ,  $H_B^{**}$  are given by

- (i)  $E_B^*(1) = \frac{H_B^* - H_B^{*(Min)}}{H_B^{*(Max)} - H_B^{*(Min)}}$      $E_B^*(1) = \frac{H_B^*}{H_B^{*(Max)}}$   
 $E_B^*(2) = \frac{H_B^*}{H_B^{*(Max)} + H_B^{*(Min)}} / 2$      $E_B^*(2) = \frac{2H_B^*}{H_B^{*(Max)}}$
- (ii)  $E_B^{**}(1) = \frac{H_B^{**} - H_B^{**(Min)}}{H_B^{**(Max)} - H_B^{**(Min)}}$      $E_B^{**}(1) = \frac{H_B^{**}}{H_B^{**(Max)}}$
- (iii)  $E_B^{**}(2) = \frac{H_B^{**}}{H_B^{**(Max)} + H_B^{**(Min)}} / 2$      $E_B^{**}(2) = \frac{2H_B^{**}}{H_B^{**(Max)}}$

Further, the modified Dominance Indices based on  $E_B^*$  and  $E_B^{**}$  are given by

- (i)  $D_B^*(1) = [1 - E_B^*(1)]$
- (ii)  $D_B^*(2) = [1 - E_B^*(2)]$
- (iii)  $D_B^{**}(1) = [1 - E_B^{**}(1)]$
- (iv)  $D_B^{**}(2) = [1 - E_B^{**}(2)]$ .

Remarks: The modified Biological Evenness Indices based on Shannon’s and Simpson’s Diversity Indices can be obtained as follows:

**I Based on Shannon’s Diversity Index**

- (i)  $E_S(1) = \frac{H_S}{H_S(\text{Max})}$
- (ii)  $E_S(2) = \frac{H_S - H(\text{Min})}{H_S(\text{Max}) - H(\text{Min})}$  and
- (iii)  $E_S(3) = \frac{H_S}{\frac{H_S(\text{Max}) + H(\text{Min})}{2}}$ .

**II Based on Simpson’s Diversity Index**

- (i)  $E_{\text{Sim}}(1) = \frac{H_{\text{Sim}}}{H_{\text{Sim}}(\text{Max})}$
- (ii)  $E_{\text{Sim}}(2) = \frac{H_{\text{Sim}} - H_{\text{Sim}}(\text{Min})}{H_{\text{Sim}}(\text{Max}) - H_{\text{Sim}}(\text{Min})}$
- $$E_{\text{Sim}}(3) = \frac{H_{\text{Sim}}}{\frac{[H_{\text{Sim}}(\text{Max}) + H_{\text{Sim}}(\text{Min})]}{2}}$$

**5 Relationships Among Goodness of Fit Test Statistics (Likelihood Ratio and Chi-Square Statistics) and Biological Diversity Indices**

Usually, one may quantify the Evenness of the species (categories) abundance distribution  $\hat{P}$  by comparing it to the perfectly Even abundance distribution given by  $P^*$ . Thus,  $P^*$  may be considered as a ‘Null model’ under the null hypothesis. To assess the departure from the perfect Evenness null model based on the observed abundance  $\hat{p} = (p_1, p_2, \dots, p_k)$ , one may use two types of goodness of fit test statistics namely,

- (i) The likelihood Ratio Test Statistic:  $LR = \Lambda = 2 \sum_{i=1}^k n_i \log\left(\frac{n_i k}{n}\right)$  and

(ii) The chi-square test statistic:  $\chi^2 = \sum_{i=1}^k \left[ \frac{(n_i - n/k)^2}{n/k} \right]$

Under information theory, instead of comparing the values of these test statistics with the critical values of Chi-distribution one may use these values as measures of the degree of departure of the true species (categories) abundance distribution from perfect Evenness abundance distribution. Suppose that the Evenness contained in two types of Kull-back-Leibler divergence measures namely,

(i)  $H_{KL}^{(1)} = \sum_{i=1}^k P_i \log \left( \frac{P_i}{P_i^*} \right)$

(ii)  $H_{KL}^{(2)} = \sum_{i=1}^k P_i \left[ \frac{p_i}{p_i^*} - 1 \right]$

Consider,

(i) 
$$\Lambda = 2 \sum_{i=1}^k n_i \log \left( \frac{n_i k}{n} \right) = 2n \left[ \sum_{i=1}^k (n_i/n) \log \left( \frac{n_i}{n} \right) + \log k \right]$$

$$= 2n \left[ \log k - \left( - \sum_{i=1}^k p_i \log p_i \right) \right]$$

$$\Lambda = 2n \left[ H_S^{(Max)} - H_S \right] \quad \left[ \because H_S^{(Max)} = \log k \right]$$

$$\frac{\Lambda}{2n} = \left[ H_S^{(Max)} - H_S \right]$$

where  $H_S$  is Shannon’s Diversity Index and  $H_S^{(Max)}$  is the Maximum possible value of  $H_S$ .

Here,  $H_S$  is sample based estimates for Shannon’s Diversity Index. Further,  $H_S^{(Max)} = \log k$  is the maximum value of  $H_S$ , if the species (categories) abundance distribution is completely even (i.e.,  $p = p^*$ )

(ii) 
$$\chi^2 = \sum_{i=1}^k \left[ \frac{(n_i - n/k)^2}{n/k} \right] = n \left[ \sum_{i=1}^k (n_i/n)^2 k - 1 \right]$$

$$= n \left[ \sum_{i=1}^k p_i^2 k - 1 \right] = n \left[ H_{Sim} k - 1 \right]$$

$$= n \left[ \frac{H_{Sim}}{1/k} - 1 \right] = n \left[ \frac{H_{Sim}}{H_{Sim}^{(Max)}} - 1 \right] \quad \left[ \because H_{Sim}^{(Max)} = \frac{1}{k} \right]$$

$$\frac{\chi^2}{n} = \left[ \frac{H_{Sim}}{H_{Sim}^{(Max)}} - 1 \right] \text{ or } \frac{\chi^2}{n} = k \left[ H_{Sim} - H_{Sim}^{(Max)} \right]$$

$$\text{or } \frac{\chi^2}{n} = -k \left[ H_{Sim}^{(Max)} - H_{Sim} \right] \text{ or } \frac{\chi^2}{2n} = \frac{-k}{2} \left[ H_{Sim}^{(Max)} - H_{Sim} \right]$$

where  $H_{Sim}$  is Simpson’s Diversity Index and  $H_{Sim}^{(Max)}$  is the maximum possible value of  $H_{Sim}$ .

Here,  $H_{Sim}$  is sample based estimate for Simpson's Diversity Index. Further,  $H_S^{(Max)} = \log k$  is the Maximum value of  $H_{Sim}$ , if the species (categories) abundance distribution is completely even (i.e.,  $p = p^*$ )

### Remarks

1. The test statistics  $\frac{A}{2n}$  and  $\frac{\chi^2}{2n}$  can be considered as sample estimates of  $H_{KL}^{(1)}$  and  $H_{KL}^{(2)}$  respectively
2. It can be easily shown that  $\frac{A}{2n}$  and  $\frac{\chi^2}{2n}$  are asymptotically unbiased Maximum Likelihood estimated of  $H_{KL}^{(1)}$  and  $H_{KL}^{(2)}$  respectively.

## 6 Conclusions

The term 'BIOMETRY' was conceived between 1892 and 1901 by Karl Pearson, along with British Journal the name "BIOMETRIKA", first published in 1901. At present, Biometrics or Biostatistics has a wide coverage of applications and contributions not only from Health, Medicine and Nutrition but also from fields Agriculture, Genetics, Biology, Biochemistry, Demography, Epidemiology, Anthropology and many others. According to 'VEDIC STATISTICS' the subject statistics was known as 'VRIKSHA VIDYA' or 'AKSHA VIDYA' originated during the period of Ancient Indian Emperor 'RUTHU PARNA', the very first Statistician (Especially Biostatistician) in counting the total number of leaves in 'KALI VRIKSHA' (Terminalia Bellerica Tree: Botanical Name) by using the concept of "SAMPLING". Emperor 'RUTHU PARNA' exchanged his 'VRIKSHA VIDYA' with 'ASWA VIDYA' by Emperor 'NALA'. At present (year 2018) the age of subject statistics is 10, 36, 17, 118 years (traced from ITHIHASAMS; Quoted by Prof. PAGADALA BALA SIDDAMUNI, PROFESSOR IN STATISTICS, Sri Venkateswara University, Tirupati along with PADMA VIBHUSHAN Prof. C. R. Rao on 24th February, 2004 in connection with 85th Birthday celebrations of Prof. C. R. Rao held at MAHATI AUDITORIUM, TIRUPATI, ANDHRA PRADESH, INDIA. Generally, for nominal scale biological data, Arithmetic mean, Median, Standard Deviation etc., may not be served as reference to study the Dispersion or Variation in the Nominal scale data. Instead, one may study the concepts of 'Diversity' and 'Evenness' of the distribution of observations among categories. In the present research work, an attempt has been made by developing some new Biometrical methods to measure important biological concepts of Diversity and Evenness involved in the 'Biometrical Genetics'. Firstly, various measures of Biological Diversity, the distribution of indistinguishable observations among categories have been described for Biometrics. Secondly, some important Biological Evenness measures have been presented in the present study. Thirdly, newly modified Biological Diversity and Evenness indices have been developed by using Brillouin, measures based on Sterling's approximation. Further,

modified Biological Evenness indices have been proposed by using Shannon's and Simpson's Diversity indices. Finally, various relationships among Goodness-of-Fit statistics namely, Likelihood Ratio, Chi-Square statistics and Biological Diversity and Evenness indices for nominal scale data have been established in the present research work.

## References

1. C. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**,379–423 (1948).  
A.L. Sheldon, Equitability indices: dependence on the species count. *Ecology* **50**, 466–467 (1969)
2. B. Smith, J.B. Wilson, A Consumer's guide to evenness indices. *Oikos* **76**, 70–82 (1996)
3. L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1962), pp. 351
4. E. Pielou, The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13**, 131–144 (1966)
5. E.H. Simpson, Measurement of diversity. *Nature* **163**, 688 (1949)
6. A. Renyi, *Probability Theory*. (North Holland, Amsterdam, 1970)
7. M.O. Hill, Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–431 (1973)
8. C. Heip, P. Engels, Comparing species diversity and evenness indices. *J. Marine Biol. Assoc.* **54**,559–63 (1974)
9. G.P. Patil, C. Taille, An overview of diversity. ecological diversity in theory and practice, in *International Cooperative Publishing House*, Fairland, MD, pp. 3–27 (1979)
10. G.P. Patil, C. Taithe, Diversity as a concept and its measurement. *J. Am. Stat. Assoc.* **77**, 548–567 (1982)
11. H.J. Zar, *Biostatistical Analysis*, 5th edn. Prentice Hall, Upper Saddle River, NJ (2009)
12. C. Tsallies, Entropic non extensivity: a possible measure of complexity. *Chaos, Solitons Fractals* **13**, 371–391 (2002)
13. A. Camargo, Must dominance increase with the number of subordinate species in competitive interactions. *J. Theor. Biol.* **161**, 537–542 (1993)
14. K.O. Bowman, K. Hutcheson, E.P. Odum, L.R. Shenton, Comments on the distribution of indices of diversity. *Stat. Ecol.* **3**, 315–366 (1971)
15. G. Madhusudan, Biometrical methods for quantitative genetics. Unpublished Ph.D. thesis in statistics (S.V. University, Tirupati, 2018)

# Data Analysis on Biopsies of Breast Cancer Tumors Data Using Data Science



K. Hemalatha, K. Hema, and V. Deepika

**Abstract** Data Science is a blend of various tools, Algorithms and Machine Learning Principles to extract knowledge from the structured and unstructured data. It is an important scientific field which uses statistics, computing science and intelligence Science to uncover the insights and trends in the data. Data Science is playing a crucial role in medical field. In this study, biopsy of breast cancer patient's data is considered to uncover the insights of the data that can help to improve the breast cancer diagnosis accuracy. The biopsy data is preprocessed to manage the missing values and the data is visualized through various graphs to demonstrate the data distribution. Supervised learning classifiers such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Feed-forward Neural Network (FNN) are used to classify the data. The performance of each classifier is compared and the best classifier that can classify the biopsy data efficiently is analyzed.

**Keywords** Data science · Breast cancer · Data visualization · Supervised learning · Support vector machine · K-Nearest Neighbors · Feed-forward neural network

## 1 Introduction

Data plays a vital role in every field. Now-a-days most of the data is from different sectors, channels, and platforms including cell phones, social media, e-commerce sites, healthcare surveys, and Internet searches. The increased amount of data opened the door to a new field of study called Big data. Big data deals with massive data. However the ever-increasing data is mostly unstructured. This leads to parsing of data for effective decision making. Hence, Data Science emerges as an important scientific

---

K. Hemalatha (✉) · K. Hema · V. Deepika  
Department of MCA, Sri Venkateswara College of Engineering (SVCE), Tirupati, India  
e-mail: [hemalathakulala@gmail.com](mailto:hemalathakulala@gmail.com)

K. Hema  
e-mail: [goldenhema@gmail.com](mailto:goldenhema@gmail.com)

V. Deepika  
e-mail: [deepika.v1@svcolleges.edu.in](mailto:deepika.v1@svcolleges.edu.in)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_40](https://doi.org/10.1007/978-3-030-46939-9_40)

field. Data Science integrates tools from multi disciplines to gather data set, process it and derives insights from the dataset. It extracts meaningful information and interpret it for decision making process. Data Science combines various disciplines such as mining, statistics, machine learning analytics and programming [1]. Applications of Data Science explores various domains such as Fraud and Risk Detection, Healthcare, Speech Recognition, Augmented Reality and etc. Among them Healthcare is most significant domain.

Breast cancer is one of the deadliest cancers that occurs in women and a second leading cause of cancer-induced death [2]. In many cases, breast cancer can be cured with the diagnosis using diagnostic mammogram and breast ultrasound. If cancer cannot be eliminated then the patients should have a biopsy. Biopsy is the test in which cells from the suspicious are of the breast will be collected and it will be examined through microscope. Biopsy is the significant breast cancer diagnostic procedure that can definitely determine if the suspicious area is cancerous [3].

Hence, the data related to the biopsies of breast cancer patients is considered in this study. For the better diagnosis of breast cancer the insights of the data and automated classification is necessary. The development of an automated diagnosis system is essential to help physician in the decision making about the cancer. Human and visual errors in traditional diagnosis methods may lead to inaccurate diagnosis. Early diagnosis and accurate disease detection is vital to reduce the death percentage among breast cancer patients. Therefore, significant Data Science techniques such as Data Cleaning, Data Visualization, and Classification are necessary.

## ***1.1 Research Objective***

The objective of this study is to uncover the insights of the data, exploring the data for better analysis and to provide a comparative study on the significant supervised learning classifiers (Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Feed-forward Neural Network (FNN)) on the biopsy data of breast cancer patients.

## ***1.2 Research Scope***

Initially the data related to the biopsies of breast cancer will be preprocessed to manage the missing values in the dataset. The data will be visualized to explore the uncover insights which can help to improve the diagnosis accuracy. Finally, three models are developed to classify the data into benign and malignant using supervised learning techniques. The developed models are evaluated and the best model among them is identified by comparing the classification performance measures of each classifier.



## 2 Materials and Methods

### 2.1 Dataset

Breast cancer database obtained from the University of Wisconsin Hospitals [4] is considered in this study. Biopsies of breast tumours for 699 patients was assessed in this dataset. The dataset consists of 699 samples and 11 features. The 11 features are listed in the Table 1.

Among 11 features 10 features are the predictor variables and the last feature (i.e. class feature) is the response variable. The feature bare nuclei contains 16 missing values. Names such as ID, V1, V2, V3, V4, V5, V6, V7, V8, V9 are assigned to the 1–10 features respectively for the purpose of feasibility to do the analysis.

### 2.2 Data Wrangling

Data wrangling is the process of cleaning and merging disorganized and complex data sets for easy access and analysis. Data wrangling is the set of actions that allows you to move from raw data to refined data, or from refined data to optimized, production data. Data forms the backbone of any data analytics. Regarding data, there are many things to go wrong in the construction, arrangement, formatting, spellings, duplication, extra spaces, and so on. To perform the data analytics properly various data wrangling techniques necessary so that the data will be ready for analysis. Generally Data scientist spend most of their time to clean and manipulate the data and spend less time to analyze the data [5].

**Data Imputation** Data Imputation is one of the steps in data wrangling. Imputation is the process in which missing data will be replaced with substituted values. Missing

**Table 1** List of features in the Biopsy dataset

S. No.	Name used in R-studio	Feature name
1	ID	Sample code number
2	V1	Clump thickness
3	V2	Uniformity of cell size
4	V3	Uniformity of cell shape
5	V4	Marginal adhesion
6	V5	Single epithelial cell size
7	V6	Bare nuclei
8	V7	Bland chromatin
9	V8	Normal nucleoli
10	V9	Mitoses
11	Class	Class

data may lead to the problems such as: (i) introduce ample quantity of bias (ii) handling and analysis of the data may become more difficult, (iii) reduce the efficiency. Various types of Imputation techniques such as Imputation using Mean/Median values, Imputation using more frequent or Zero values, Imputation using KNN and etc. are available to replace the missing data. Among the above said Imputation techniques Imputation using Mean/Median values is described in this section because this imputation technique is used in this study [6].

**Imputation Using Mean/Median Value** In this technique the mean/median of non-missing values will be calculated. Then the missing values will be replaced with the calculated mean/median values within each column separately. This technique is applicable to numeric data only. Imputation using mean method is easiest and faster than other imputation methods. It work well for small numeric dataset.

### 3 Methodology

The proposed methodology in this study consists of phases such as

- Loading data
- Data Wrangling
- Data Exploration and Visualization
- Classification Modeling
- Model Evaluation.

In the first phase biopsy data on breast cancer patients which available on web is loaded into R-studio. In the dataset the feature bare nuclei consists of 16 missing values which may affect the classification efficiency. Hence, in the Data Wrangling phase the 16 missing values are replaced using the imputation using mean/median technique. The mean of the column bare nuclei is calculated and the missing values are replaced by the mean. After Data Wrangling phase the biopsy data is explored and visualized. Data related to all features in the dataset is visualized using histograms in Fig. 1.

Class wise data (benign and malignant) is visualized using histograms in Fig. 1. The data with benign class is represented in pink color and the data with malignant class is represented in blue color. From the figure it is known that the samples with benign class are more than the sample with malignant class in the dataset.

To see how the predictors are related to each other, bivariate relationships are considered and they are plotted in Fig. 2.

There are few variables that are correlated. Often the features with highest correlation value may provide redundant information. Hence highly correlated feature that is ID is eliminated to reduce the predictive bias.

In Classification modelling phase, before applying the various classifiers to the data the input dataset is spilt into training and test datasets. 70% of the data is



separated as training set and the remaining 30% of the data is separated as test set. After the splitting of data significant supervised learning techniques such SVM, KNN and FNN are applied to the data and three classification models are build. In Model evaluations phase the performance of the proposed three models is compared and the best model is evaluated.

### 4 Results and Analysis

In this section, the results and analysis for SVM, KNN and FNN are provided. Further, a comparative analysis of the proposed three models is also presented. All the experiments are conducted in R-studio 3.6.0. SVM is trained using the input data with parameters: kernel function is radial,  $\gamma$  value equal to 0.090 (i.e. 1/dimensions) and coefficient value equal to zero. All the parameters used are the default parameters. KNN and FFN classifiers are also trained with the training data and the predictive accuracies are calculated for the test data.

The confusion matrix obtained in the testing phase for SVM, KNN and FFN are represented in the Table 2.

From the obtained confusion matrices the predictive accuracy, sensitivity and specificity of the three models is calculated. The results are presents in Table 3.

The three models predicted the data with higher accuracies. Among the three model FFN predicted the data with 100% accuracy, specificity and sensitivity. Whereas SVM and KNN predicted the data with the accuracies 97.13% and 99.52% respectively. After FFN, KNN predicted the data with highest accuracy. SVM prediction accuracy, Sensitivity are less when compared to the other two models. But the Specificity of the all the models are same i.e. 100%.

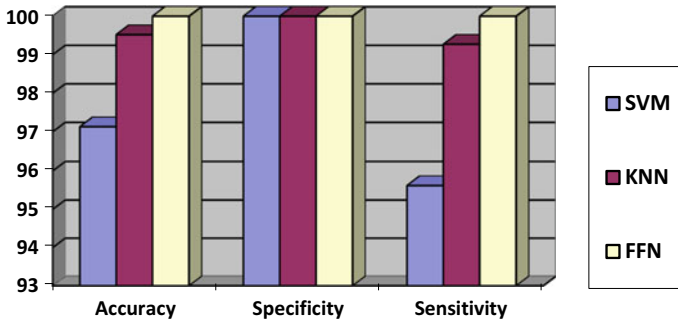
The prediction results are presented graphically in Fig. 3.

**Table 2** Confusion matrices of SVM, KNN and FFN

Prediction	SVM		KNN		FFN	
	Benign	Malignant	Benign	Malignant	Benign	Malignant
Benign	131	0	136	0	137	0
Malignant	6	72	1	72	0	72

**Table 3** Prediction results of SVM, KNN and FFN

S. No	Model	Accuracy (%)	Specificity (%)	Sensitivity (%)
1	SVM	97.13	100	95.62
2	KNN	99.52	100	99.27
3	FFN	100	100	100



**Fig. 3** Comparison of classifiers performance

From the graph it can be observed that the Specificity of all the models is same. Accuracy, Specificity and Sensitivity of the FNN model is highest than the other models. Hence FNN model is the best fit for the biopsy data on breast cancer patients.

## 5 Conclusions

Data Science, one of the emerging techniques provides efficient solutions to the Healthcare problems. Automated breast cancer diagnosis using biopsy data is necessary to avoid the risk related to it. Hence, data science techniques such as Data Exploration, Data Visualization, Data Wrangling and Classification Modelling are used in this study. The biopsy data on breast cancer patients is pre-processed, the insights in the data is explored and visualized. Finally three classification models using the supervised learning techniques such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Feed-forward Neural Network (FNN). The predictive performance of the three models is evaluated. All the three models performed well and predicted data with higher accuracies. But among them FNN classified the data with 100% Accuracy. Hence, FNN is concluded as the best fit for the biopsy data of breast cancer patients.

## References

1. <https://www.investopedia.com/terms/d/data-science.asp>. Retrieved on 23/09/2019
2. M.R. Ataollahi, J. Sharifi, M.R. Paknahad, A. Paknahad, Breast cancer and associated factors: a review. *J Med Life*. **8**(4), 6–11 (2015)
3. <https://www.nationalbreastcancer.org/breast-cancer-biopsy>. Retrieved on 25/09/2019
4. P.M. Murphy, D.W. Aha, UCI repository of machine learning databases. (Machine-readable data repository). University of California, Department of Information and Computer Science, Irvine, CA (1992)

5. J. Han, J. Pei, M. Kamber, *Data Mining: Concepts and Techniques*, in 3rd edn. The morgan kaufmann series in data management systems (Morgan Kaufmann, Burlington, 2011). ISBN: 978-0123814791
6. <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>. Retrieved on 27/09/2019

# A Comparison of Multi Support Vector Machine Performance with Popular Decomposition Strategies on Alzheimer's Data



R. M. Mallika, K. Usha Rani, and K. Hemalatha

**Abstract** In Medical field, classifying the images into more than two diagnostic classes is still remaining as a challenge. Particularly in diagnosis of Alzheimer's the brain MRI images should be classified into more classes for effective diagnosis. Multi Support Vector Machine (MSVM) is advancement to the standard SVM which can able to deal multi-class classification problem efficiently. MSVM is successfully applied in the fields of Text categorization, Medical Image classification, Handwriting Recognition, Protein Structure Prediction, etc. MSVM uses numerous approaches such as Directed Acyclic Graph, One-vs-One and One-vs-All etc., to classify the data into multi classes. Among them "One vs One" and the "One vs All" are the important decomposition strategies. Hence, the both decomposition strategies are applied separately with MSVM in this study to classify the Alzheimer's disease data. The performance of MSVM with these two strategies is compared and the best decomposition strategy is identified.

**Keywords** Support vector machine · Multi-class classification · Brain MRI images · Feature extraction · Texture features

## 1 Introduction

Multi-Class Classification is one in all the classical issues in Machine Learning [1]. In Multi-Class Classification problem each input should be classified into one of the finite set of categories (classes) [2]. Few applications of Multi-Class Classification includes Part-of-speech Tagging, Handwritten Optical Character Recognition, Image

---

R. M. Mallika (✉) · K. Usha Rani  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [mallika.521@gmail.com](mailto:mallika.521@gmail.com)

K. Usha Rani  
e-mail: [usharanikuruba@yahoo.co.in](mailto:usharanikuruba@yahoo.co.in)

K. Hemalatha  
Department of MCA, Sri Venkateswara College of Engineering, Tirupati, India  
e-mail: [hemalathakulala@gmail.com](mailto:hemalathakulala@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_41](https://doi.org/10.1007/978-3-030-46939-9_41)

Categorization, etc. The classification of Brain MRI Images is one of the problems of Multi-Class classification in which images should be classified into more than two diagnostic classes. The diagnosis of Alzheimer's disease is one of the difficulties of Multi-Class Classification of Brain MRI Images. While diagnosing the Alzheimer's the MRI Brain Images should be classified into three different stages, they are (i) Alzheimer's Disease(AD) (ii) Mild Cognitive Impairment(MCI) (iii) Normal Control (NC). Diagnosing Alzheimer's at an early stage is necessary to take preventive care which can help to mitigate risk factors [3]. The prognosis of Alzheimer's disease can be done with the aid of efficient classification of Brain MRI Images.

SVM plays a vital role in Medical Diagnosis. SVM is linear model for classification and regression problems [3]. In SVM a hyperplane which splits the data points into classes which takes the data as an input and outputs a line that separated those classes if possible. Hyperplane is closed to the data points (support vectors) from each of the classes. The power of SVM is that it allows inference at individual level instead of group level [3]. SVM are inherently binary classifier. But the diagnosis of Alzheimer's disease includes the classification of MRI images of Brain into AD, MCI, NC. Therefore SVM alone is not sufficient. To overcome this problem MSVM is considered here. MSVM decompose the multi-class problem into several binary classification problems [4]. Hence, here we are going to study the performance of MSVM on MRI images by considering the axial projections which are proved for Alzheimer's disease classification.

Here, the OASIS database [5] which is publicly available is considered for Brain MRI image classification. The MRI images of Brain are preprocessed and the Texture features of MRI images are extracted. Texture is usually used feature in the evaluation and interpretation of images. To categorize the texture by colors or pixel intensities of an image. Texture analysis is used in various applications of Computer image analysis for reduction, segmentation and classification of images based on spatial distribution of intensity levels [6]. The main objective of this examine is to evaluate the overall MSVM performance with two popular decomposition strategies "One vs One" and "One vs All" to categorize the MRI images of Brain into AD, MCI and NC.

This paper is structured as follows. In Second section, the associated works are shown. In Third section, the Methods and Materials are presented along with the proposed method implementation. In Fourth section, the experiment results are described. The paper is concluded in Sect. 5.

## 2 Literature Survey

Few studies related to Support Vector Machine and Multi-class Classification are reviewed in this section for the Alzheimer's disease diagnosis.

Anthony et al. [2] has presented an overview of SVM Classification of MRI Images. In this study, the two methods are used i.e., One-Against-One (OAO) and One-Against-All (OAA) with accuracy evaluation.



Few studies related to Support Vector Machine and Multi-class Classification are reviewed in this section for the Alzheimer's disease diagnosis.

Anthony et al. [2] has presented an overview of SVM Classification of MRI Images. In this study, the two methods are used i.e., One-Against-One (OAO) and One-Against-All (OAA) with accuracy evaluation.

Jhosi et al. [7] has obtained overview of statistical structure analysis based tumor segmentation to extract the texture features using GLCM technique. ANN and Fuzzy c-means approaches are considered for the categorization of disease diagnosis.

Benabdeslem et al. has presented a novel approach Dendo gram based SVM makes use of taxonomy of classes and decompose a multiclass problem into set of two-class problems by considering the databases like "Iris", "Glass" and "Letter". High classification accuracy can be obtained through DSVM constantly.

Xian long Liu et al. has proposed a Multiclass SVM model.

Iris, Dermatology, sat image, svm guide data sets are considered for experimentation in this study.

Ahuja et al. [6] has presented the various Multiclass Classification approaches One-vs-One, One-vs-All and Directed Acyclic Graph and represents fitness of Support Vector Machines in multi class classification.

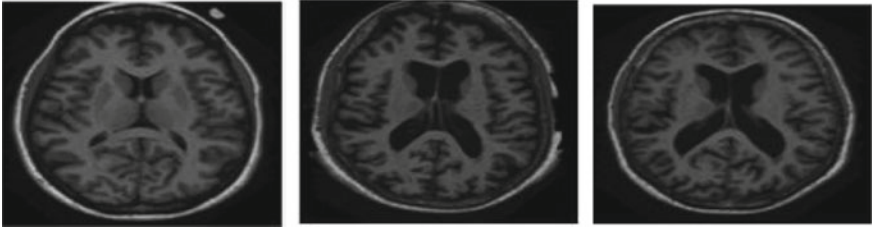
Mallika et al. [8] has proposed A Fuzzy-Based Expert System to Diagnose Alzheimer's Disease. In this study, MRI images of three projections such as axial, coronal and sagittal from OASIS database are considered. To classify the MRI images into AD, MCI, and Non AD subjects Fuzzy Inference System is developed. The classification performance of FIS is more for axial projection with accuracy of 86.53% and sensitivity of 92.73% than coronal, sagittal projections.

By the motivation from above studies Multi-Class SVM is considered for experimentation in this study to diagnose the Alzheimer's.

## 3 Materials and Methods

### 3.1 Dataset

In this study, Open Access Series of Imaging Studies (OASIS) database consists of a cross sectional collection of 416 T1 weighted Brain Magnetic Resonance Imaging (MRI) scanned Images is used for experimentation [9]. Dataset consists of MRI Images of Brain with three types of projections such as Axial, Coronal and Sagittal. Among these three projections axial projections based images are successfully used for better diagnosis of Alzheimer's disease. Hence, in this study among three projections only axial projections based images are considered for experimentation [8, 10]. Sample axial projections of Brain MRI images are shown in Fig. 1.



**Fig. 1** Sample brain MRI images of axial projection

### 3.2 Feature Extraction

The conversion of an image into its set of features is known as Feature Extraction. Suitable features of image are extracted from the MRI images for classification work. Extracting a good feature set for classification is a very tedious task. Some of the Feature Extraction techniques such as Linear Discriminant Analysis, Principal Component Analysis, Independent Component Analysis [11], Clustering methods etc.

The Texture features associated to Brain MRI images are extracted in feature extraction phase by using Gray Level Co-occurrence Matrix (GLCM). Extracting second order statistical texture features can be done using GLCM method. The number of gray levels,  $G$ , of the image is equal to the no. of rows and columns of GLCM matrix [12]. The GLCM element  $P(i, j | \Delta x, \Delta y)$  is the absolute frequency with the two pixels, separated by a pixel distance  $(\Delta x, \Delta y)$ , occur within the given area, with two intensity levels  $i, j$ . The matrix element  $P(i, j | d, \theta)$  comprises the second order statistical probability values for changes between gray levels 'i' and 'j' at a specific displacement distance 'd' at a particular angle  $(\theta)$ . By large number of intensity levels  $G$  implies filling a lot of temporary data, i.e., a  $G \times G$  matrix for each combination of  $(\Delta x, \Delta y)$  or  $(d, \theta)$ . This matrix is very sensitive to the size of the texture samples in which they are approximate due to its high dimensionality. So, the numbers of gray levels are frequently reduced. To extract texture features from image the best proven method is GLCM [13]. Some of the important features are shown in Table 1. The GLCM used to generate statistical texture parameters such as Mean, Correlation, homogeneity, Energy, Entropy, standard deviation, variance, inertia, skewness, etc.

### 3.3 Multi-class Support Vector Machine (MSVM)

Support Vector Machine can suitable only for binary classification. Nowadays large amount of data is necessary for classification and the data should be classified into two or more classes. So, Binary Classification alone is not sufficient. In such cases the need for multi-class classification arises. Hence, the SVM is enhanced by several researchers to solve Multi-class Classification problems. In Binary SVM, the hyper

**Table 1** Texture features extracted

S. No.	Texture feature	S. No.	Texture feature
1	Mean	9	Geometric compactness (GC)
2	Standard deviation (SD)	10	Texture mean (TM)
3	Contrast	11	Texture global mean (TGM)
4	Energy	12	Texture standard deviation (TSD)
5	Correlation	13	Texture smoothness (TS)
6	Homogeneity	14	Texture uniformity (TU)
7	Geometric area (GA)	15	Texture entropy (TE)
8	Geometric perimeter (GP)	16	Texture skewness (TSK)
		17	Texture correlation (TC)

planeis dividing the data into two groups. By considering this, SVM is extended as MSVM by numerous methods like Directed acyclic graph, One-vs-One and One-vs-All, etc., Among these approaches One-vs- One, One-vs-All are the most popular approaches.

### 3.4 One-vs-All Decomposition Strategy

In One-vs-All approach N different binary classifiers should be build to classify the data into N classes. For ith classifier all the points in the class are considered as positive samples and which are not in the class are considered as negative samples. In this strategy rather than predicting a class label the base classifier produces confidence score for its decision. The decision will be taken by applying all classifiers to an unknown sample x and predicting the category label k that the corresponding confidence score is highest. The unknown sample x belongs to the class k if that class has largest decision function value. The decision function is presented by the Formula 1 [6].

$$(y) = \arg \max_k f_k(x) \tag{1}$$

where  $f_k$  is a list of classifiers for  $k \in \{1, \dots, K\}$ .

### 3.5 One-vs-One Decomposition Strategy

The Strategy One-vs-One overcomes the limitation existed using One-vs-All approach by training more binary SVM models. In this approach one binary SVM model for each two classes will be trained. Therefore totally  $c(c-1)/2$  models will

be trained for  $c$  classes. The classification will be done based on the voting strategy.  $c(c-1)/2$  binary classifiers are applied to an unknown sample and that sample will be classified as a particular class that contain highest votes [14].

One-vs-All approach suitable one classifier consistent with one class wherein as One-vs-One approach fit one classifier consistent with pair of classes. One-vs-All is the most commonly used method because of its benefit interpretability. One-vs-One does not provide better results when the sample size increases [11].

### 3.6 Methodology

Methodology using MSVM with two popular decomposition approaches such as “One-vs-One” and “One-vs-All” is proposed in this study. The proposed Methodology includes three stages: Preprocessing, Feature Extraction and Classification. The work flow of the proposed methodology is presented in Fig. 2. The MRI Images are preprocessed to eliminate the noise and to improve the image by highlighting the edges in the first stage. In Second stage, GLCM is used to extract the seventeen texture features listed in Table 1 from axial projections of Brain MRI Images. To classify the data into multiple classes by using MSVM classifier in third stage. MSVM with One-vs-One approach and MSVM with One-vs-All approach are experimented separately to classify the Alzheimer’s data.

It is difficult to show all the extracted feature values in one table. Hence, for convenience the extracted feature values of 10 brain MRI images samples are presented in Table 2. Using One-vs-One approach, for each two classes one binary SVM model

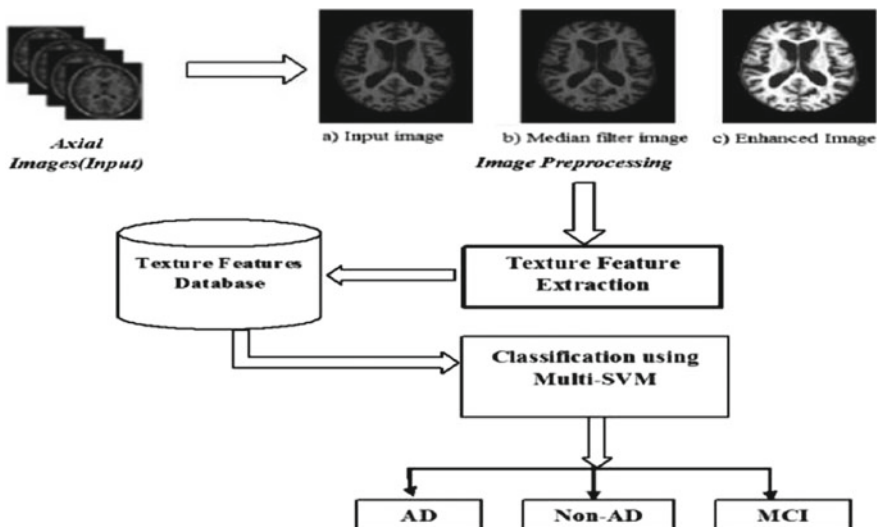


Fig. 2 Workflow of proposed methodology

**Table 2** Extracted texture features of 10 brain MRI images

Texture features	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6	Image 7	Image 8	Image 9	Image 10
Mean	36.43	36.75	39.51	43.15	40.36	40.43	39.42	43.23	40.43	31.65
SD	40.93	34.61	45.11	35.95	46.36	34.42	45.97	39.13	48.67	38.38
Contrast	0.1	0.85	0.51	0.18	0.1	0.17	0.13	0.22	0.15	0.12
Energy	0.32	0.21	0.3	0.27	0.35	0.27	0.34	0.25	0.34	0.38
Correlation	0.96	0.55	0.85	0.91	0.97	0.9	0.96	0.92	0.96	0.95
Homogeneity	0.95	0.79	0.88	0.93	0.95	0.92	0.94	0.9	0.93	0.94
GA	18,310	22,984	18,016	31,979	18,310	29,495	18,310	29,014	18,310	18,310
GP	473	1837	921	518	473	567	473	522	473	473
GC	12.22	146.82	47.08	8.39	12.22	10.9	12.22	9.39	12.22	12.22
TM	72.83	57.61	80.28	49.4	80.69	50.17	78.81	54.54	80.84	63.29
TGM	36.43	36.75	39.51	43.15	40.36	40.43	39.42	43.23	40.43	31.65
TSD	26.42	26	29.34	34.22	32.31	31.33	33.48	36.26	38.34	30.7
TS	0.00143	0.001400	0.001000	0.000853	0.000957	0.001018	0.000891	0.000760	0.000680	0.001060
TU	0.33	0.17	0.41	0.11	0.41	0.12	0.4	0.15	0.44	0.27
TE	-172.86	-146.95	-186.24	-133.11	-186.8	-134.3	-182.95	-142.93	-185.86	-152.14
TSK	-0.37	0.52	0.03	0.29	-0.54	0.33	-0.36	0.38	-0.28	-0.18
TC	5999.81	4019.29	7302.88	3609.32	7551.65	3497.18	7329.75	4288.23	8003.18	4945.6

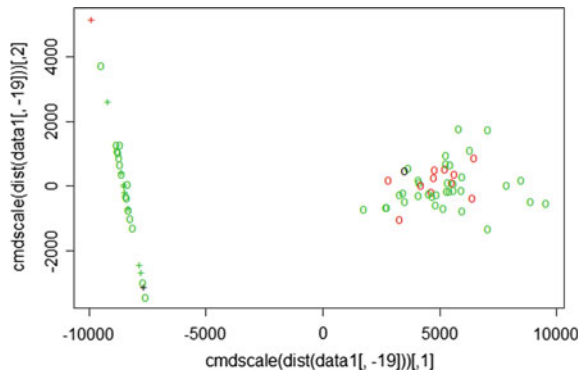
is trained. Totally 3 models are trained for 3 classes such as AD, Non-AD and MCI. Based on the voting strategy 3 binary classifiers are implemented to the data and the data is classified into a specific class based on the highest votes. In One-vs-All approach all the three binary classifiers are applied to the data and predicted the category label based on the corresponding high confidence score.

The score is calculated using the formula 1. Similarly entire data is classified into three classes such as AD, Non-AD and MCI.

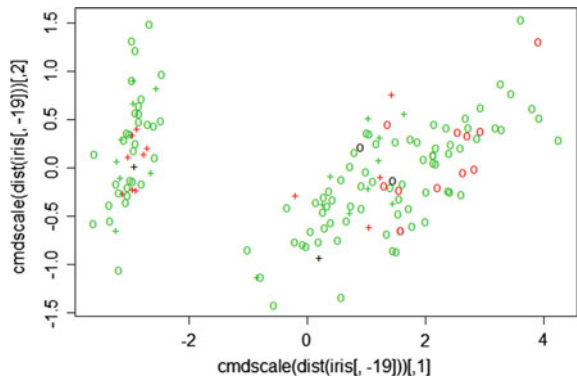
### 4 Results and Analysis

In this study MSVM is experimented separately with two popular decomposition techniques like “One-vs-One” and “One-vs-All”. For the classification purpose the texture features dataset is extracted in feature extraction phase. MSVM classifies the data into three classes such as AD, MCI and NC using the above said decomposition techniques. The classification is visualized in the below mention Figs. 3 and 4.

**Fig. 3** Visualization of classes predicted using MSVM with One-vs-All approach



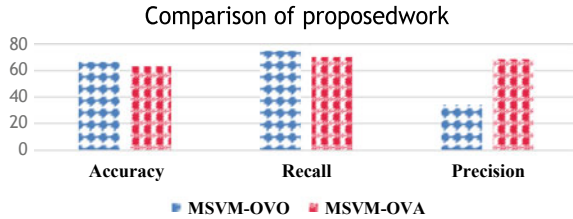
**Fig. 4** Visualization of classes predicted using MSVM with One-vs-One approach



**Table 3** Comparison of MSVM performance with One-vs-One and One-vs-All approaches

Method	Accuracy (%)	Recall (%)	Precision (%)
MSVM with One-vs-One	<b>66.7</b>	<b>75</b>	34.4
MSVM with One-vs-All	63.51	70.45	<b>68.88</b>

**Fig. 5** Comparison of performance of MSVM with OVO and OVA



From the Figs. 3 and 4 it can be observed that the data samples are represented in three colors such as black, red and green. These colors are associated with the classes of AD, MCI and NC respectively. Along with colors two shapes such as O and + are presented in the both figures. The shape O represents the target data and the shape + represents the predicted data using MSVM. The performance of classification measures like Accuracy, Recall and Precision are evaluated for the two models. The calculated classification performance measures are presented in Table 3.

MSVM classifier with One-vs-One (OVO) approach classified the data with 66.7% of accuracy, 75% of Recall and 34.4% of Precision. Whereas MSVM classifier with One-vs-All (OVA) approach classified the data with 63.51% of accuracy, 70.45% Recall and 68.88% Precision. From the results it is observed that MSVM with One-vs-One approach classified the data with higher accuracy than One-vs-All approach. Recall value is also high for One-vs-One approach than One-vs-All. But precision value of One-vs-One approach is less than One- vs-All approach. The comparison of MSVM classification measures is represented diagrammatically in Fig. 5.

MSVM with OVO classified the data with highest accuracy than MSVM with OVA approach. The Recall value is also high for MSVM with OVO. But the precision value in MSVM with OVO is less than the MSVM with OVA approach.

Classifier with high recall value and low precision is preferable for medical dataset. Recall express the ability to find all relevant cases where as precision identifies the relevant data points [13]. Hence, MSVM with OVO is the optimal model that perform better than MSVM with OVA.

## 5 Conclusion

SVMs are popular binary classifier. From literature SVM is not appropriate to categorize the brain MRI Images into multiple classes. MSVM is the development to the usual SVM which could able to deal Multi-Class classification problem efficiently. Problems related to Medical field are almost Multi-Class Classification problems. Multi-class classification of medical data is necessary in the prediction of diseases effectively. Hence, in this study we considered the Multi-class classification of Brain MRI images into Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI) and Normal Control (NC). Multi Support Vector Machine (MSVM), a significant Multi-class classifier with two popular decomposition techniques "One-vs-One (OVO)" and "One-vs-All (OVA)" is used for the experimentation. The MRI Images of Brain are preprocessed and the significant texture features are extracted. The extracted features dataset is used for the classification. MSVM with OVO and MSVM with OVA model are experimented. The classification performance of MSVM with the above said decomposition approaches are evaluated using the classification measures such as Accuracy, Recall and Precision. MSM with OVO performed better than MSVM with OVA. Hence, MSVM with OVO will be considered for further studies. The performance of MSVM classification could be enhanced using the Feature Selection Techniques in future.

## References

1. V. Vapnik, *Statistical Learning Theory*, vol. 1. Wiley (1998)
2. G. Anthony, H. Greg, M. Tshilidzi, Classification of images using support vector machines. Department of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, Private Bag X3, Wits, South Africa (2005)
3. P. Vemuri, C.R. Jack, Role of structural MRI in Alzheimer's disease. *Alzheimer's Res. Ther.* 2, 23 (2010)
4. S. Huang, N. Cai, P.P. Pacheco, S Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics.* 15(1):41–51 (2017). 0.21873/cgp.20063. Epub
5. Open Access Series of Imaging Studies available. [www.oasis-brains.org](http://www.oasis-brains.org), <http://www.oasrains.org/app/template/Index.vm>
6. Y. Ahuja, S.K. Yadav, Multiclass classification and support vector machine. *Global J. Comput. Sci. Technol.* 12(11), (2012)
7. J. Joshi, A.C. Phadke, Feature extraction and texture classification in MRI
8. R.M. Mallika, K. UshaRani, K. Hemalatha, A fuzzy-based expert system to diagnose alzheimer's disease. *Internet of Things and Personalized Healthcare Systems*. Springer Briefs in Applied Sciences and Technology. Springer, Singapore (2019)
9. OASIS brain Alzheimer dataset. [www.oasis-brains.org/](http://www.oasis-brains.org/)
10. P. Keserwani et al., Classification of alzheimer disease using gabor texture feature of hippocampus region. *Int. J. Image, Graph. Signal Proc.* 6, 13–20, (2016). Published Online June 2016 in MECS
11. <http://www.statsoft.com/textbook/support-vector-machines>
12. P. Mohanaiah, P. Sathyanarayana, L. GuruKumar, Texture feature extraction using GLCM approach. *Int. J. Sci. Res. Publ* 3(5), 2 (2013). ISSN: 2250-3153



13. S. Mary Joans, J. Sandhiya, A genetic algorithm based feature selection for classification of brain MRI scan images using random forest classifier. *Int. J. Adv. Eng. Res. Sci. (IJAERS)*, **4**(5), (2017). <https://dx.doi.org/10.22161/ijaers.4.5.20>. ISSN: 2349-6495(P)2456-1908(O)
14. C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **13**
15. H. Helen F. Zhang, *Multiclass support vector machine* (2017)

# Synthesis, Evaluation and in Silico Studies of 4-N, N-Dimethylamino and 4-Carboxy Chalcones as Promising Antinociceptive Agents



Shaheen Begum, S. K. Arifa Begum, A. Mallika, and K. Bharathi

**Abstract** In the present work, a series of substituted 4-N, N-dimethylamino and 4-carboxy chalcones were synthesized using Claisen-Schmidt condensation and characterized by spectral data. The antinociceptive activity was studied in mice using different models. The results indicated that in the 4-dimethylamino series, compound 4c, bearing 3,4-dimethoxy substituent has shown good peripheral antinociceptive activity. Among the 4-carboxy series, 5c and 5i were found to be potent both in central and peripheral pain models. In dimethylamino chalcones, compounds 4b and 4c containing 4-methoxy and 3, 4-dimethoxy groups showed good binding affinity towards iNOS, while compound 4f bearing 4-chloro substituent exhibited good binding affinity towards COX-2 enzyme. Interestingly, in the series of carboxy chalcones compound 5e containing 4-fluoro substitution demonstrated good affinity towards both iNOS and COX-2 enzymes.

**Keywords** Dimethylamino chalcones · 4-Carboxy chalcones · Antinociceptive activity · Molecular docking

---

S. Begum (✉) · K. Bharathi

Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam (Women's University), Tirupati, Andhra Pradesh, India  
e-mail: [Shaheen.pharmchem@gmail.com](mailto:Shaheen.pharmchem@gmail.com)

S. K. Arifa Begum

Avanathi Institute of Pharmaceutical Sciences, Jawaharlal Nehru Technological University Hyderabad (JNTUH), Gunthapally, Hyderabad, Telangana, India

A. Mallika

Department of Medicinal Chemistry, National Institute of Pharmaceutical Education and Research Hyderabad, Balanagar, Hyderabad, India

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_42](https://doi.org/10.1007/978-3-030-46939-9_42)

## 1 Introduction

Chalcones have gained considerable research interest as they exhibit a variety of biological activities. Chalcones exhibit antinociceptive activity probably by inhibiting iNOS and COX-2 enzymes. Presence of substituent groups on ring A or B of chalcones showed profound effect on the analgesic activity of these molecules. Literature revealed that the introduction of acidic moiety such as carboxyl group on ring B of naturally occurring chalcone, flavokawain B significantly increased its activity both in peripheral and central antinociceptive pain models [1, 2].

Based on the reports indicating the imperative role of inflammatory mediators (NO & PGE<sub>2</sub>) in pain and perception and the potentiality of dimethylamino substituted chalcones as NO and PGE<sub>2</sub> inhibitors, it was emphasized that substitution with dimethylamino group may exert significant contribution to antinociceptive activity [3]. In view of the above findings, we planned to synthesize a series of 4-N, N-dimethylamino or 4-carboxy chalcones with various substituent groups on 3rd and 4th positions of the other phenyl ring and screened for analgesic activities using mice. Docking studies were performed with COX-2 and iNOS to predict their probable binding mode.

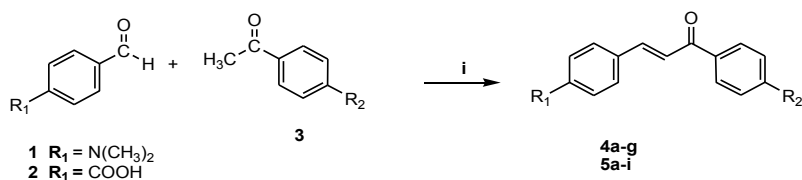
## 2 Experimental

### 2.1 Materials and Methods

Bruker Advance spectrometer was used for spectral analysis.

#### General procedure for the synthesis of chalcones (4a-g and 5a-i) [3]

Equimolar (5 mmol) amounts of acetophenone and substituted aldehyde (5 mmol) were dissolved in 30 ml methanol. The reaction mixture was stirred by slowly adding NaOH (50%, 5 ml) at room temperature. Product was obtained once the reaction contents were neutralized with HCl (1 N) (scheme 1) [4].



**4a-g** R<sub>1</sub> = N(CH<sub>3</sub>)<sub>2</sub> =R<sub>2</sub>= H, 4-OCH<sub>3</sub>, 3,4-(OCH<sub>3</sub>)<sub>2</sub>, 4-CH<sub>3</sub>, 4-F, 4-Cl, 4-NO<sub>2</sub>

**5a-i** R<sub>1</sub> = COOH = R<sub>2</sub>= H, 4-OCH<sub>3</sub>, 3,4-(OCH<sub>3</sub>)<sub>2</sub>, 4-CH<sub>3</sub>, 4-F, 4-Cl, 4-NO<sub>2</sub>, 4-C(CH<sub>3</sub>)<sub>3</sub>, 3,4,5-(OCH<sub>3</sub>)<sub>3</sub>

**Scheme 1** Synthesis of 4-N,N-dimethylamino and 4-carboxy chalcones. Reagents and reaction conditions: (i) 50% NaOH, CH<sub>3</sub>OH, r.t., 1-24 h

**Spectral (proton NMR) details of the compounds**

**4a) 3-(4-(dimethylamino) phenyl)-1-phenylprop-2-en-1-one.** 8.37–6.78 (m, 1H, Ar-H, CH=CH) 3.03 (s, 6H, N (CH<sub>3</sub>)<sub>2</sub>). **4b) 3-(4-(dimethylamino) phenyl)-1-(4-methoxyphenyl) prop-2-en-1-one.** 8.01 (d, 2H, Ar-H) 7.71 (d, 2H, CH=CH) 7.65 (d, 2H, *J* = 15.7), 7.36 (d, 2H, Ar-H) 6.76 (d, 2H, Ar-H) 3.01 (s, 6H, N (CH<sub>3</sub>)<sub>2</sub>) 3.8 (s, 3H, OCH<sub>3</sub>) **4c) 1-(3,4-dimethoxyphenyl)-3-(4-(dimethylamino)phenyl) prop-2-en-1-one** 8.04 (s, 3H, Ar-H) 7.72 (d, H $\alpha$ , 15.6 Hz) 7.43 (d, H $\beta$ , *J* = 15.7 Hz), 6.99–6.86 (m, 4H, Ar-H) 3.01 (s, 6H, N (CH<sub>3</sub>)<sub>2</sub>) 3.92 (s, 6H, (OCH<sub>3</sub>)<sub>2</sub>) **4d) 3-(4-(dimethylamino) phenyl)-1-p-tolylprop-2-en-1-one.** 8.02 (d, 2H, Ar-H) 7.70–7.36 (m, 6H, Ar-H) 6.75(d, 2H, CH=CH) 3.00 (s, 6H, N (CH<sub>3</sub>)<sub>2</sub>). 2.50 (s, 3H, CH<sub>3</sub>). **4e) 3-(4-(dimethylamino) phenyl)-1-(4-fluorophenyl) prop-2-en-1-one.** 7.97(s, 2H, Ar-H), 7.71(d H $\alpha$ , *J* = 15.6), 7.34(d, H  $\beta$ , *J* = 15.6 Hz) 6.97–6.84(m, 5H, Ar-H), 3.03 (s, 6H, N(CH<sub>3</sub>)<sub>2</sub>) **4f) 1-(4-chlorophenyl)-3-(4-(dimethylamino phenyl) prop-2-en-1-one.** 7.95–6.58 (m, 9H, Ar-H), 7.09(d, 2H, CH=CH) 2.80 (s, 6H, N (CH<sub>3</sub>)<sub>2</sub>) **4g) 3-(4-(dimethylamino) phenyl)-1-(4-nitrophenyl) prop-2-en-1-one.** 8.37–7.77 (m, 8H, Ar-H) 6.77 (d, 2H, CH=CH) 3.03 (s, 6H, N (CH<sub>3</sub>)<sub>2</sub>).

**5a) 4-(3-oxo-3-phenylprop-1-enyl) benzoic acid:** 13.01 (s, 1H, COOH) 8.01-7.30 (m, 6H, Ar-H) 7.07 (d, H $\alpha$ , *J* = 16.2), 6.31 (d, H $\beta$ , *J* = 16.2 Hz), 6.21(s, 2H, Ar-H) **5b) 4-(3-(4-methoxyphenyl)-3-oxoprop-1-enyl)benzoic acid** 13.00 (s, 1H, COOH) 8.02–7.79 (m, 5H, Ar-H) 7.29 (d, H $\alpha$ , *J* = 16.2 Hz) 7.05 (d, H $\beta$ , *J* = 16.2 Hz), 6.30(s, 2H, Ar-H) 3.83 (s, 3H, OCH<sub>3</sub>) **5c) 4-(3-(3,4-dimethoxyphenyl)-3-oxoprop-1-enyl) benzoic acid** 13.01 (s, H,-COOH), 8.04–7.80 (m, 5H, Ar-H) 7.29 (d, H $\alpha$ , *J* = 16.4 Hz), 7.07 (d, H $\beta$ , *J* = 16 Hz) 6.30 (s, 2H, Ar-H) 3.83 (s, 5H, (OCH<sub>3</sub>)<sub>2</sub>) 3.71(s, 1H, OCH<sub>3</sub>) **5d) (4-3-oxo-3-p-tolylprop-1-enyl) benzoic acid** <sup>1</sup>H-NMR: 13.05 (s, H,-COOH), 8.11–7.85 (m, 8H, Ar-H) 7.79(d, H $\alpha$ , *J* = 16.4 Hz), 7.50 (d, H $\beta$  *J* = 16 Hz), 2.43 (s, 3H, CH<sub>3</sub>) **5e) 4-(3-(4-fluorophenyl)-3-oxoprop-1-enyl) benzoic acid:** 13.01 (s, 1H, COOH) 8.01–7.30 (m, 6H, Ar-H) 7.07 (d, H $\alpha$ , *J* = 16.2), 6.31 (d, H $\beta$ , *J* = 16.2 Hz), 6.21(s, 2H, Ar-H) **5f) 4-(3-(4-chlorophenyl)-3-oxoprop-1-enyl) benzoic acid:** 13.01 (s, 1H, COOH) 8.01–7.30 (m, 6H, Ar-H) 7.07 (d, H $\alpha$ , *J* = 16.2), 6.31 (d, H $\beta$ , *J* = 16.2 Hz), 6.21(s, 2H, Ar-H) **5g) 4-3-(4-nitrophenyl)-3-oxoprop-1-enyl) benzoic acid:** 13.29 (s, H,-COOH), 8.01–7.30 (s, 6H, Ar-H) 7.07 (d, H $\alpha$ , *J* = 16.2 Hz), 6.31 (d, H $\beta$ , *J* = 16.2 Hz), 6.21(s, 2H, Ar-H) **5h) 4-(3-(4-tert-butylphenyl)-3-oxoprop-1-enyl) benzoic acid:** 13.08 (s, H,-COOH), 8.12–7.93 (m, 4H, Ar-H) 7.80 (d, H $\alpha$ , *J* = 16.09 Hz) 7.75(d, H $\beta$ , *J* = 16.09 Hz) 7.50 (d, 2H, Ar-H) 1.26 (s, 9H, (CH<sub>3</sub>)<sub>3</sub>) **5i) 4-(3-(3,4,5-trimethoxyphenyl)-3-oxoprop-1-enyl)benzoic acid:** 13.13 (s, H,-COOH), 8.04–7.81 (m, 5H, Ar-H) 7.30 (d, H $\alpha$ , *J* = 16.4 Hz), 7.09 (d, H $\beta$  *J* = 16 Hz), 6.32 (s, 2H, Ar-H) 3.84 (s, 3H, OCH<sub>3</sub>) 3.72(s, 6H, (OCH<sub>3</sub>)<sub>2</sub>).

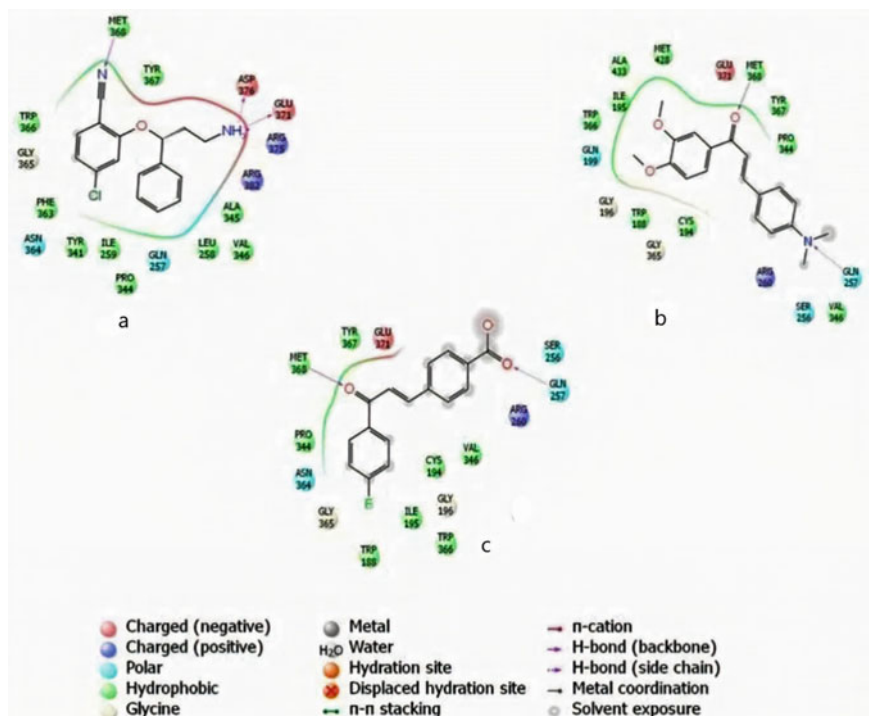


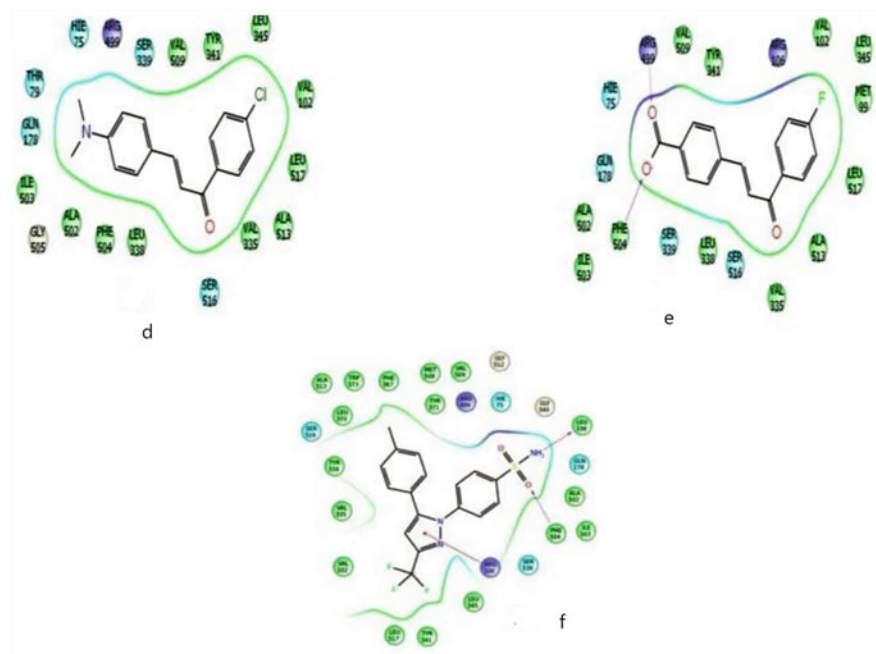
Fig. 1 Binding orientation of a standard ligand b 4b c 5e in the binding site of iNOS enzyme

## 2.2 Antinociceptive Studies

Male mice (Swiss albino, 25–30 g) were used (1677/PO/a/12/IAEC/44). Four analgesic models; acetic acid induced writhing test [5], hot-plate [6], formalin method [7], tail-immersion test [8] were used (Figs. 1 and 2; Tables 1 and 2).

## 2.3 Docking Studies

Glide (Schrödinger suite 2013) module was used and 3D-structures of COX-2 (3LN1) and iNOS (2Y37) were taken from PDB. Grid co-ordinates were based on the crystal ligand of PDB. For COX-2 were X: 65.245; Y: -45.3789; Z: 45.4229 and for iNOS X: 20.78937; Y: -68.949104; Z: 32.876694 respectively.



**Fig. 2** Binding orientation of chalcones **d** **e** **f** celecoxib in the binding site of COX-2 enzyme

## 2.4 Results and Discussion

A series of chalcones were synthesized based on the observations that introduction of carboxyl group on ring B of chalcones, improved the antinociceptive activity [2]. Efficient pharmacokinetic features and the ability to participate in electrostatic interactions and hydrogen bonds are the key characteristics of carboxyl functional group. In previous studies, dimethylamino chalcones showed potent iNOS inhibitory activity [3]. Several derivatives were also evaluated for their antiinflammatory activity and presence of 3,4-dimethoxy moiety was proved to be promising for the activity. These findings suggest the probable contribution of dimethylamino moiety to antinociceptive activity [3, 9].

Keeping in view the usefulness of the two functional moieties, we prepared some novel chalcones containing the acidic (4-carboxy) or basic substituents (4-dimethyl amino) to study the influence of these features on antinociceptive activity.

### Acetic acid induced writhing assay

Among the 4-N,N-dimethylamino chalcones (**4a-g**), compound **4b**, possessing methoxy substitution at the 4th position was found to be effective (72.59%), indicating the importance of strong electron releasing effect of methoxy group for the

**Table 1** Antinociceptive activities of dimethylamino (4a-g) and carboxy chalcones (5a-i) in different central and peripheral antinociceptive models

Compound No.	R	Tail immersion test (latency period in sec)	Hot-plate test (latency period in sec)	A. A induced writhing (PP)	Formalin test	
					Number of lickings (0–5 min)	Number of lickings (15–30 min)
Control		3.167 ± 0.1667*	2.500 ± 0.2236*			
Standard		11.55 ± 0.22 <sup>a</sup> *	8.33 ± 0.2108 <sup>b</sup> *	82.46 <sup>c</sup>	82.98 <sup>d</sup>	87.50 <sup>d</sup>
4a	4-H	4.0 ± 0.8912*	3.5 ± 0.55*	18.2	26.90	–
4b	4-OCH <sub>3</sub>	7.2 ± 0.7543*	5.7 ± 0.51*	72.59	47.24	48.45
4c	3,4-(OCH <sub>3</sub> ) <sub>2</sub>	9.2 ± 0.7521*	3.5 ± 0.54*	78.06	45.26	67.19
4d	4-CH <sub>3</sub>	5.5 ± 0.5412*	4.3 ± 0.51*	24.77	24.16	–
4e	4-F	7.7 ± 0.8112*	5.5 ± 1.51*	63.15	37.91	32.83
4f	4-Cl	6.7 ± 0.5156*	3.5 ± 0.54*	56.80	26.47	29.70
4g	4-NO <sub>2</sub>	7.2 ± 0.8112*	3.5 ± 0.54*	48.90	31.54	–
5a	4-H	6.7 ± 0.5156	3.5 ± 0.55	33.81	24.16	21.27
5b	4-OCH <sub>3</sub>	7.0 ± 1.26	5.7 ± 1.63	46.71	39.90	30.64
5c	3,4-(OCH <sub>3</sub> ) <sub>2</sub>	8.5 ± 1.51	7.0 ± 1.26*	70.65	42.30	23.14
5d	4-CH <sub>3</sub>	3.1 ± 1.26	4.0 ± 0.89	27.38	0.70	–
5e	4-F	7.7 ± .81	5.5 ± 1.51*	57.76	27.46	51.26
5f	4-Cl	5.2 ± 0.75*	5.5 ± 1.63	34.21	20.87	21.27
5g	4-NO <sub>2</sub>	5.2 ± 0.75	6.0 ± 1.41*	43.68	17.57	26.89
5h	4-C(CH <sub>3</sub> ) <sub>3</sub>	3.4 ± 0.81	3.5 ± 0.55	20.76	1.30	–
5i	3,4,5-(OCH <sub>3</sub> ) <sub>3</sub>	9.5 ± 0.54*	7.0 ± 1.26*	74.73	57.13	53.13

Values are expressed as mean ± SEM, n = 6; Values represent one-way analysis of variance (ANOVA) followed by Dunnett's test

PP Percentage inhibition

\**p* < 0.05 versus control

<sup>a, b</sup> tramadol

<sup>c, d</sup> aspirin

peripheral antinociceptive activity. 3, 4-dimethoxylated derivative, **4c** displayed highest activity in acetic acid writhing assay (78.06%) suggesting that there is an increase in analgesic activity with increased substitution in case of methoxylated derivatives.

In the carboxy chalcones, significant activity was noted for the compounds substituted with electron releasing groups viz., methoxy, dimethoxy and trimethoxy (**5b**, **5c** and **5i**) with 46.71%, 70.65% and 74.73% inhibition respectively suggesting the importance of methoxy substitution in eliciting peripheral antinociceptive activity. This is further supported by the facts that, the electron donating alkoxy groups on ring A of chalcones, could lower the acidity of  $\alpha$ -hydrogens which forms a more stable GSH adduct and makes the retro-Michael reaction slower. The effect of substituent is evident when the ring B contains carboxyl group at 4th position when compared to N, N-dimethylamino group indicating the importance of acidic group at this position. The combined effects i.e. electron withdrawing effect at 4th position

**Table 2** Docking scores for dimethylamino (4a-g) and carboxy chalcones (5a-i) with COX-2 and iNOS

Compound No.	R	Interaction energy (kJ mol <sup>-1</sup> )	Interacting amino acids	Interaction energy (kJ mol <sup>-1</sup> )	Interacting amino acids
4a	4-H	-7.79	-	-4.18	Glu 371
4b	4-OCH <sub>3</sub>	-7.42	Tyr 341, Arg 106	-5.93	Met368, Gln 257
4c	3,4-(OCH <sub>3</sub> ) <sub>2</sub>	-7.20	Phe 504	-5.57	Met368, Gln 257
4d	4-CH <sub>3</sub>	-7.60	Tyr 341 ( $\pi$ - $\pi$ stacking), Trp 373 ( $\pi$ - $\pi$ stacking)	-3.97	Glu 371
4e	4-F	-5.90	-	-5.58	Met368, Gln 257
4f	4-Cl	-8.48	Arg 106, Tyr 341( $\pi$ - $\pi$ stacking)-	-5.39	Met368
4 g	4-NO <sub>2</sub>	-5.90	-	-3.84	Met368
5a	4-H	-8.48		-5.33	Met368
5b	4-OCH <sub>3</sub>	-7.57	Arg499	-5.35	Met368
5c	3,4-(OCH <sub>3</sub> ) <sub>2</sub>	-7.15	Arg 499	-5.61	Met368, Gln257
5d	4-CH <sub>3</sub>	-8.44	Arg 499, Phe 504	-5.25	Met368, Arg 375
5e	4-F	-8.54	Arg 499, Phe 504	-5.65	Met368, Gln 257
5f	4-Cl	-8.45		-5.43	Met368, Gln257
5g	4-NO <sub>2</sub>	-7.06	-	-3.93	Met368, Arg260
5h	4-C(CH <sub>3</sub> ) <sub>3</sub>	-6.99	Tyr 341( $\pi$ - $\pi$ stacking), Tyr 371	-4.62	Met368
5i	3,4,5-(OCH <sub>3</sub> ) <sub>3</sub>	-7.31	Phe 504	-4.94	Met368, Gln257
CL-1 (COX-2)	Celecoxib	-11.87	Arg 106 ( $\pi$ -interactions), Phe 504, Leu 338 (hydrogen bonding)	-	-
CL-2 (iNOS)	2Y37_A54		-	-8.68	Met368, Glu371, Asp376



and strong electron releasing effect at positions 3, 4 and 5 are most favorable for this assay.

### Formalin-induced pain test

Among all dimethylamino chalcones, only **4b**, 3, 4-dimethoxy chalcone showed potent analgesic behavior in this assay at a dose of 50 mg/kg when given orally. Halogen containing chalcones, **4e**, **4f** and nitro derivative **4g** which exhibited moderate activity in acetic acid induced writhing assay were unable to produce any notable analgesia in the formalin-induced pain assay. In case of carboxy chalcone series, compounds which had methoxy functionality viz., **5b**, **5c** and **5i** produced analgesia both phases, and the maximum analgesia (57.13%) was observed for chalcone bearing trimethoxy groups at 3, 4 and 5 positions of ring A (**5i**), suggesting involvement of these compounds at both the central and peripheral levels.

### Tail immersion model

Derivative **5i**, with 3, 4 and 5 trimethoxy substitution demonstrated high potency (latency period  $9.5 \pm 0.54$  s). Interestingly, these methoxy chalcones **4b**, **4c**, **5b**, **5c** and **5i** which were active in this model were also effective in the acetic acid and formalin induced pain assays, indicating that these compounds have remarkable central and peripheral antinociceptive activity. The presence of fluoro or nitro substitution on dimethylamino chalcones **4e** and **4g** significantly increased the latency period with  $7.7 \pm 0.81$  and  $7.2 \pm 0.75$  s respectively. The good activity of the fluoro derivative is attributed to the desirable characteristics imparted by fluorine which can modulate pharmacokinetic and pharmacodynamic properties. Similar results were also observed in case of carboxy chalcones with the electron withdrawing groups, where in fluorine analogue **5e** showed high activity suggesting that incorporation of this group is capable of reinforcing drug-receptor interactions.

### Hot plate method

In this model, activity exerted by methoxylated derivatives was prominent compared to other substituent groups. From both series of synthesized chalcones, compounds bearing methoxy groups **4b**, **4c**, **5b**, **5c** and **5i** were capable of increasing the latency period of pain at 50 mg/kg when administered orally ( $3.5 \pm 0.55$  to  $7.0 \pm 1.26$ ).

Comparison of activities reveals that 4-carboxy substitution is favorable for both central and peripheral analgesic activities and 4-N, N-dimethylamino substitution is favorable for peripheral analgesic activity. The activity profiles of these two series of compounds highlight the important contribution of methoxy substitution at distinct positions (3, 4 and 5). In the chalcone scaffold compounds bearing methoxy substitution namely **4b**, **5b**, **5c** and **5i** showed greater effectiveness both in central and peripheral analgesic pain models.

### Docking with iNOS enzyme

Dimethylamino chalcones were well embedded in the binding site to interact with Met368 using its ketonyl oxygen. Dimethylamino group (nitrogen) was found to form a hydrogen bond with Gln257. Glide scores indicate that compound **4b**, with methoxy substitution exhibited highest interaction energy ( $-5.93$  kJ mol<sup>-1</sup>) as it can bind with Met368 and Gln257. In case of carboxy chalcones, unsubstituted chalcone (**5a**) exhibited good binding affinity ( $-5.33$  kJ mol<sup>-1</sup>) comparable to

4-methoxy derivative ( $-5.35 \text{ kJ mol}^{-1}$ ). Presence of dimethoxy substitution on carboxy chalcone (**5c**) was proved to be favorable, ( $-5.61 \text{ kJ mol}^{-1}$ ) for binding, while trimethoxylated derivative showed good in vivo activities, but in docking analysis, binding affinity ( $-4.94 \text{ kJ mol}^{-1}$ ) was decreased, indicating that bulkiness is not favorable in the active site. Substitution with halogen (**5f** and **5g**) is found to be favorable in the series of carboxy chalcones ( $-5.65$  and  $-5.43 \text{ kJ mol}^{-1}$ ). The better interaction energies of the fluoro derivatives (**4e** and **5e**) suggest fluorine as favorable substitution than chlorine for good drug receptor interaction. Results of binding studies revealed that fluorinated chalcones **4e** and **5e** have good binding energies ( $-5.58$  and  $-5.65 \text{ kJ mol}^{-1}$ ) in the active site of iNOS.

#### Docking studies with COX-2 enzyme

In the series of dimethylamino chalcones, chloro analogue (**4f**) exhibited highest binding score ( $-8.48 \text{ kJ mol}^{-1}$ ). Fluoro and nitro substitution (**4e** and **4f**) led to the compounds with decreased binding efficiency ( $-5.90$  and  $-5.90 \text{ kJ mol}^{-1}$ ) whereas methylation and monomethoxylation increased binding capability ( $-7.60$  and  $-7.42 \text{ kJ mol}^{-1}$ ). Carboxy chalcones with their carbonyl oxygen (C=O) and negatively charged oxygen were able to establish hydrogen bonds with critical amino acids such as Arg499 and Phe504 respectively. Hydrophobic interactions were also observed in the vicinity of this enzyme active site with the amino acids Val335, Ala513, Leu517, Val102 and Tyr341 with these chalcones. Highest binding affinity with the interaction energy of  $-8.54 \text{ kJ mol}^{-1}$  was observed for 4-fluoro derivative, **5e**, suggesting that electronic modulation and small covalent radius of fluorine could facilitate docking with their drug receptors. Methyl substitution is more favorable ( $-8.44 \text{ kJ mol}^{-1}$ ) in comparison to mono and dimethoxy substitution ( $-7.57$  and  $-7.15 \text{ kJ mol}^{-1}$ ).

### 3 Conclusion

Sixteen chalcone derivatives were synthesized and characterized by spectroscopy. Among the screened compounds, compound **5i** was found to be the most promising chalcone with the potential to be developed as antinociceptive agent. Further studies using suitable ELISA kits will enhance better understanding of these compounds on these enzymes and will help in the development of potent drug candidates.

### References

1. R. Correa, A.S. Pereira, D. Buffon, L. Santos, V.C. Filho, A.R.S. Santos, R.J. Nunes, Antinociceptive Properties of chalcones. Structure-activity relationships. *Archiv der Pharmazie*. **334**, 332–334 (2001)
2. F. Campos-Buzzi, J.P. Campos, P.P. Tonini, R. Correa, R.A. Yunes, P. Boeck, V. Cechinel-Filho, Antinociceptive effects of synthetic chalcones obtained from xanthoxyline. *Arch. Pharm.* **339**, 361–365 (2006)

3. J. Rojas, J.N. Dominguez, J.E. Charris, G. Lobo, M. Paya, M.L. Ferrandiz, Synthesis and inhibitory activity of dimethylamino-chalcone derivatives on the induction of nitric oxide synthase. *Eur J. Med. Chem.* **37**, 699–705 (2002)
4. R. Li, G.L. Kenyon, F.E. Cohen, X. Chen, B. Gong, J.N. Dominguez, E. Davidson, G. Kurzban, R.E. Miller, E.O. Nuzum, *J. Med. Chem.* **38**, 5031–5037 (1995)
5. H.D.J. Collier, L.C. Dinnin, C.A. Johnson, C. Schneider, The abdominal constriction response and its suppression by analgesic drugs in the mouse. *Br. J. Pharmacol.* **32**, 295–310 (1968)
6. N.B. Eddy, D. Leimback, Synthetic analgesic-II, Dithienyl butenyl and dithienyl butylamines. *J. Pharmacol Exp. Ther.* **107**, 385–393 (1953)
7. A.T. Hunskaar, K. Hole, The formalin test in mice: dissociation between inflammatory and non-inflammatory pain. *Pain.* **30**, 103–104 (1987)
8. M. Zimmoman, Ethical guidelines for investigations of experimental pain in conscious animals. *Pain.* **16**, 109–110 (1983)
9. F. Campos-Buzzi, P. Padaratz, A.V. Meira, R. Correa, R.J. Nunes, Cachinel-Filho, V:4'-Acetamidochalcone derivatives as potential antinociceptive agents. *Molecules* **12**, 896–906 (2007)

# In Silico Analysis for Detection of Glucose Transport-2 Inhibitors from Seagrass



Mathakala Vani, Narem Ritesh Siddhartha Reddy,  
and Palempalli Uma Maheswari Devi

**Abstract** Inhibition of Glucose transporter-2 (GLUT2) activity during glucose absorption at intestine level is a new concept for the diabetes treatment. At present, majority of the anti-diabetic drugs are oriented towards insulin deficiency and insulin resistance to control hyperglycemia conditions. GLUT2, a sodium dependent transport protein employed a prominent role in the resorption of glucose at tubular level. Thus the inhibitors of GLUT2 target towards sugar excretory and resorption mechanisms at intestine and renal level. The present study is designed to detect the inhibitory activity of seagrass bioactive compounds against GLUT2 by in silico analysis. Molecular docking software namely Discovery studio was adopted to analyse the docking potential of the ligands. The compounds derived from seagrass showed remarkable binding energies than the standards Glibenclamide and Metformin. The results demonstrate the anti-diabetic potential of seagrass bioactive leads through the inhibition of GLUT2 transport protein.

**Keywords** Glucose transporter-2 (GLUT2) · Seagrass · CDOCKER · Diabetes mellitus (DM)

## 1 Introduction

Diabetes is a heterogeneous life style disease resulting from insufficient secretion of insulin or insulin dysfunction. DM is characterized hyperglycemia due to alterations in the metabolism of carbohydrates, lipids and proteins [1]. Globally, the occurrence of DM is on rise, particularly type 2 diabetes mellitus is predominant form and 90% of the cases reported to be T2DM with characteristic feature of insulin resistance [2].

---

M. Vani · P. Uma Maheswari Devi (✉)  
Department of Applied Microbiology and Biochemistry, Sri Padmavati Mahila Visvavidyalayam,  
Tirupati 517502, India  
e-mail: [umadevi66@gmail.com](mailto:umadevi66@gmail.com)

N. Ritesh Siddhartha Reddy  
Sri Venkateswara Medical College, Tirupati 517507, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_43](https://doi.org/10.1007/978-3-030-46939-9_43)

Glucose homeostasis in the body depends upon the glucose output and utilization between hepatic cells and insulin-dependent tissues such as adipocytes and skeletal muscle [3, 4]. About 80% of glucose from circulation was cleared by skeletal muscle during glucose and insulin infusion, while adipose tissues account for much less. Thus, on the whole, skeletal muscle plays a prominent role in balancing controlled regulation of sugar under insulin-stimulated conditions.

Currently, several drugs are in use to decrease blood glucose levels by targeting the enzymes involved in glucose metabolism for the prevention and management of type 2 diabetes [5]. All these hypoglycemic agents are intertwined with side effects like gastrointestinal and kidney disorders. Thus, there is a need to select the natural products for the regulation of glucose metabolism by targeting glucose transport proteins. The inhibitors of Glucose transporter 2 (GLT2) are molecules with a mechanism different from the presently available therapies.

Glucose transporters are the proteins, which are important in the facilitative diffusion of glucose across the plasma membranes [6–8]. Among the glucose transporters [9, 10], transporter 2 is highly precise in tissue-specific transport of glucose, particularly export and import of glucose in the hepatocytes and absorption and re-absorption of glucose from intestinal brush borders and kidney tubule cells respectively [11, 12]. Due to its bidirectional role, GLUT2 is an important target in the treatment of type 2 diabetes.

Current analysis is supported with *in silico* protein-ligand interaction. Based on the significance of glucose transport protein -2 in the management of diabetes, the present research carried out to identify the natural ligands as inhibitors of GLUT-2 [13].

## 2 Materials and Methods

### 2.1 *In Silico Methods*

The homology modeling software (i.e. MODELER) is used to predict GLUT2 full-length structures. Next, the predicted models' geometry validation is carried out using PROCHECK and ProSa web server. The X-ray structure of GLUT2 was generated by using ProSa web server.

### 2.2 *Homology Modeling of GLUT2 Structure*

To predict full-length structures of human glucose transporter type GLUT 2, we have retrieved their sequence from UniProt web site (Entry ID: P11168 and P14672) and were subjected to BLASTp program to get/obtain their homologue structures from PDB. After BLASTp, the obtained PDB IDs 5EQG and 4PYP were selected as templates for modeling of GLUT 2 structure. The selected template sequence

shares 53% identity and 99% query coverage with GLUT2 sequence. Next, the template sequences were aligned with their respective target GLUT2 sequences using CLUSTAL W module. Further, the aligned sequences were used to build model structures using MODELERE program (<https://salilab.org/modeller/>). Total 10 models were generated for each target for GLUT2. Subsequently, the generated models were scored and ranked based on their molpdf scores. Further, their geometry reliability was verified using PROCHECK web servers. Finally, the best reliable structure of GLUT2 was chosen for further study.

### ***2.3 Validation of Protein Structure***

The structure of Glucose transport-2 protein was assessed by considering the parameters of back bone chain, side chain residues, phi bonds, psi bonds, root mean square distance and geometry of planar groups [14].

### ***2.4 Preparation of Glucose Transport Protein***

The modeled protein structure of GLUT2 was prepared using 'Prepare Protein' module of Discovery Studio (DS) software suit. While preparation, missing side chains atoms were filled, and bond length and dihedral angles were fixed. Further, the protein residue ionization and protonation states were derived at pH 7.4. Then, the structures were optimized by employing CHARMM force field. The binding sites of the prepared protein structure were defined using 'Edit Binding Site' option available on the tool bar of DS under receptor-ligand interaction site. The radius of GLUT2 was calculated 24 and 25.2 Å radius.

### ***2.5 Ligand Preparation***

Ligands (plant extracts), and reference compounds (metformin, glibenclamide) were prepared using 'Prepare Ligand' module of DS. While preparation, hydrogen atoms were added and charged groups were neutralized. Further, CHARMM was adopted to measure ionization potential of ligands, and also minimal energy states of tautomers and alternative chiral centers were scored.

## 2.6 Molecular Docking

Protein-ligand interaction was carried out using CDOCKER module of Discovery studio. During docking analysis, the designed ligands were docked at the defined catalytic site of GLUT2 structure. While docking default parameter were employed. After docking, the generated ligand poses were graded based on interaction energy as per the binding energy module of DS.

## 3 Results

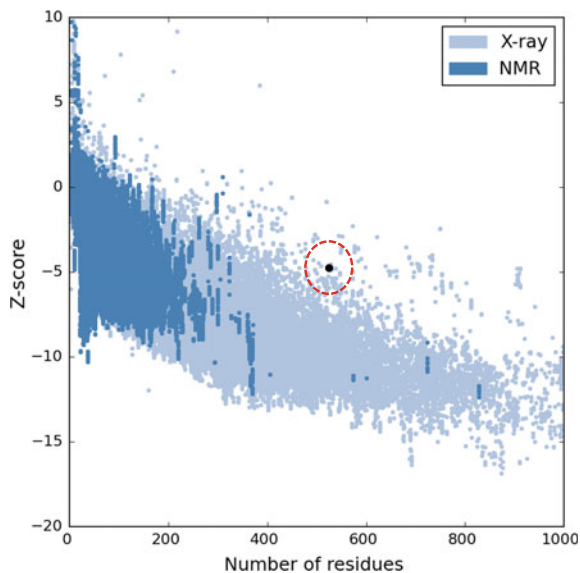
### 3.1 Homology Modeling of GLUT2 Structure

The modeled structure of GLUT2 appears like a typical barrel shape transporters. Further, the model validation results shows that modeled structures are having reliable structural geometry and were used for docking study our plant extracts. The predicted structures of GLUT2 were represented in Fig. 1. The ProSa web server results revealed that the generated model GLUT2 quality belongs to X-ray structure (Fig. 2).

**Fig. 1** Ribbon representation of modeled structures of GLUT2



**Fig. 2** Representation of model quality using ProSa web server of GLUT



### 3.2 Structural Assessment of the Protein

Further, the Ramchandran Plot for all residue types are given in Fig. 3, the PROCHEK results revealed that GLUT2 structure contains 93.3% residues in favorable region, 4.658% residues in additional allowed region, 1.5% residues in generously allowed region and 0.4% residues in disallowed region. Therefore, the PROCHEK results of GLUT2 were represented in Fig. 3.

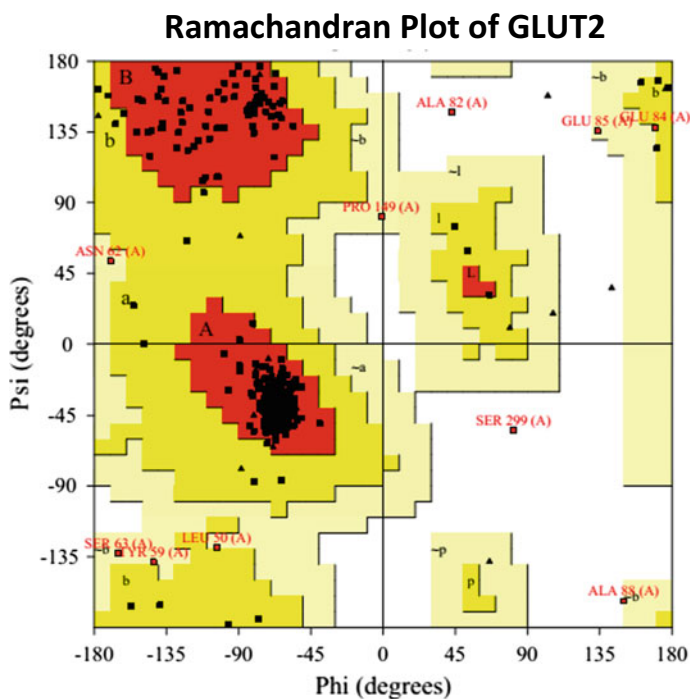
### 3.3 Molecular Docking of Phytochemicals with GLUT2

To determine the mode of binding and molecular mechanism of protein-ligand interaction the prepared ligands were docked with GLUT2 as described in methodology. After docking analysis the graded protein-ligand interactions were used to measure the binding affinity of different ligands with GLUT2. The obtained docking and binding energies were represented in Table 1 and Fig. 4.

### 3.4 Binding Mode of Cis-9 Octadecenoic with GLUT2

Binding mode analysis of Cis-9 Octadecenoic acid with GLUT2 revealed that residues Ser169, Pro173, Met174, His192, Gln193, Ile196, Gln314, Phe411, Gly416, Trp420, Ala424, Leu436 and Phe492 were involved in hydrophobic interaction; and





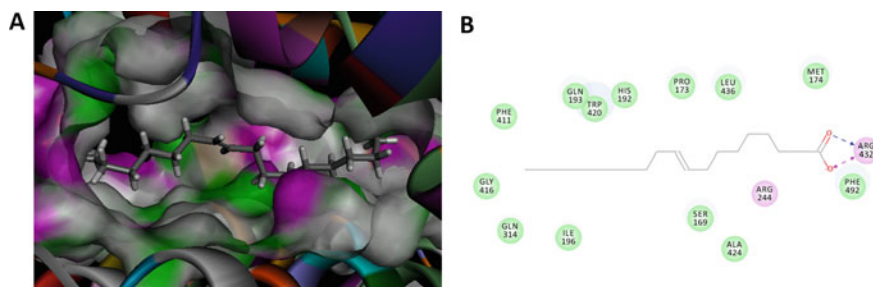
**Fig. 3** Schematic representation of Ramachandran plot of GLUT2

**Table 1** CDOCKER and binding energy of ligands with GLUT2

Compound code	Compound name	Binding energy ( $\Delta G$ kcal/mol)	-CDOCKER energy (kcal/mol)	-CDOCKER interaction energy (kcal/mol)
4091	Metformin	-15.0935	12.3121	14.4202
3488	Glibenclamide	-36.3899	27.5009	44.2386
222284	Beta-sitosterol	-35.1168	-34.7756	40.745
8181	Methyl palmitate	-24.1437	35.7792	32.6936
1017	1,2 Benzene dicarboxylic acid	-19.1671	1.37222	16.6569
445639	Cis-9 Octadecenoic acid	-102.541	29.7898	42.2903
68108	Isocoumarin	-1.40779	7.16279	12.9789
985	Hexadecanoic acid	-2.00565	36.2622	36.7978



**Fig. 4** Representation of binding energies of the compounds



**Fig. 5** 3D and 2D binding mode of Cis-9 octadecenoic with GLUT2

residues Arg244 and Arg432 were involved in electrostatic interactions. The 3D and 2D binding mode of Cis-9 Octadecenoic with GLUT2 were represented in Fig. 5.

## 4 Discussion

In silico docking analysis is an hallmark to detect the three dimensional geometry and orientation of protein-ligand interactions. In absence of experiment structure, homology modeling could be used to predict protein structure. Here, we have used homology-modeling software (i.e. MODELER) to predict GLUT2 full length structures. Next, the predicted models geometry validation is carried out using PROCHECK and ProSa web serve.

To determine the potential antidiabetic agents, the compounds identified from GC-MS analysis were docked into the GLUT2 by using Docking module of Discovery model. Among all the compounds docked, Cis-9 Octadecenoic acid was found to be in strong association with GLUT2.

As per the literature, flavonoids suppress the glucose uptake by locking glucose transporters. Therefore, the bioactive fatty acids are proved to be the potential antidiabetic agents.

## 5 Conclusion

The natural ligand Cis-9 Octadecenoic acid exhibited minimal docking energy – 102.54 kcal/mol when compared to standard Glibenclamide (–36.38 kcal/mol) and Metformin (–2.09 kcal/mol). Thus natural compounds derived from seagrass seems to be potential GLUT-2 inhibitors and anti-hyper glycemc agents.

**Acknowledgements** Authors are thankful to CURIE laboratory.

## References

1. N.Z. Baquer, Glucose over utilization and under utilization in diabetes and effects of antidiabetic compounds. *Ann. Real Farm.* **64**, 147–180 (1998)
2. R. Agrawal, E. Tyagi, R. Shukla, C. Nath, Insulin receptor signaling in rat hippocampus: a study in STZ (ICV) induced memory deficit model. *Eur. Neuropsychopharmacol. j. euroneuro.* **21**, 261–273 (2011)
3. H. Wallberg-Henriksson, Z. Nie, J. Henriksson, Reversibility of decreased insulin-stimulated glucose transport capacity in diabetic muscle with in vitro incubation: insulin is not required. *J. Biol. Chem.* **262**, 7665–7671 (1987)
4. R.A. DeFronzo, R.C. Bonadonna, E. Ferrannini, Pathogenesis of NIDDM. A balanced overview. *Diabetes Care* **15**, 318–368 (1992)
5. C. Tang, X.A. Zhu, Specific pharmacophore model of sodiumdependent glucose co-transporter 2 (SGLT2) inhibitors. *J. Mol. Model.* **18**, 2795–2804 (2012)
6. J.S. Wu, Y.H. Peng, J.M. Wu, J.M., C.J. Hsieh, S.H. Wu, M.S. Coumar, Discovery of non-glycoside sodium-dependent glucose cotransporter 2 (SGLT2) inhibitors by ligand-based virtual screening. *J. Med. Chem.* **53**(24), 8770–8774 (2010)
7. C. Hale, M. Wang, Development of 11beta-HSD1 inhibitors for the treatment of type 2 Diabetes. *Mini. Rev. Med. Chem.* **8**, 702–710 (2008)
8. V. Derdau, T. Fey, J. Atzrodt, Synthesis of isotopically labelled SGLT inhibitors and their metabolites. *Tetrahedron* **66**(7), 1472–1482 (2010)
9. R. Rajesh, P. Naren, S. Vidyasagar, Unnikrishnan, S. Pandey, M. Varghese, Sodium glucose Co transporter 2 (SGLT2) inhibitors: a new SWORD for the treatment of type 2 diabetes mellitus. *Int. J. Phar Sci. Res.* **2**, 139–147 (2010)
10. R. Grempler, R. Augustin, S. Froehner, T. Hildebrandt, E. Simon, M. Mark, Functional characterisation of human SGLT-5 as a novel kidney-specific sodium-dependent sugar transporter. *FEBS Lett.* **586**, 248–253 (2012)
11. W. Zhanga, A. Welihinda, J. Mechanic, H. Ding, L. Zhu, Y. Lu, EGT1442, a potent and selective SGLT2 inhibitor, attenuates blood glucose and HbA1c levels in db/db mice and prolongs the survival of stroke-prone rats. *Pharmacol. Res.* **63**, 284–293 (2011)
12. I. Idris, R. Donnelly, Sodium-glucose co-transporter-2 inhibitors: an emerging new class of oral antidiabetic drug. *Diabetes obes Metab.* **11**, 79–88 (2009)
13. G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. *J. Comput. Chem.* **19**(14), 1639–1662 (1998)
14. G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963)

# Automated Diagnosis of Shoulder Pain Using Regression Algorithms



B. Triveni, P. Bhargavi, and S. Jyothi

**Abstract** Shoulder disorders are related with pain, stiffness, weakness, limited range of motion and infirmity, which in some cases may continue for several years. Shoulder disorder is considered to be one of the important musculoskeletal problem in foremost care. Every twelve months 2% of adults are expected to consult with new shoulder pain. Shoulder examination with caution will be a vital factor in making a diagnosis for the shoulder disorders. Special tests are used to assist diagnostic correctness by explicitly examining one section of the shoulder complex. Special tests are generally used to support the examiners in determining the tests that doctors can depend on and do well-versed conclusions about their medical diagnosis. In this paper Automated shoulder pain diagnosis of Rotator cuff disorders Acromioclavicular joint Separation, Instability, Impingement, Labrum tear and Biceps tendon based on special tests are predicted using Multivariable Linear Regression (MLR), Decision Tree Regressor (DT), Gradient boosting regressor (GBR), Adaboost Regressor (AdaR) and Support Vector Regressor (SVR).

**Keywords** Shoulder disorders · Diagnosis · Clinician · Special tests · Regressor

## 1 Introduction

The shoulder is a compound joint with certain range of motion, unstable, surrounded with fleshy tissue structures for steadiness. The shoulder includes bony joint, fleshy tissues, glenohumeral ligaments, coracoacromial ligament, rotator cuff muscles, scapulothoracic muscles, and the long head of the biceps [1]. Improper functioning of any one of these parts may result to shoulder problems [2]. Due to Stiffness,

---

B. Triveni (✉) · P. Bhargavi · S. Jyothi  
SPMVV, Tirupati, India  
e-mail: [btriveni16@gmail.com](mailto:btriveni16@gmail.com)

P. Bhargavi  
e-mail: [pbhargavi18@yahoo.co.in](mailto:pbhargavi18@yahoo.co.in)

S. Jyothi  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_44](https://doi.org/10.1007/978-3-030-46939-9_44)

pain, or weakness shoulder movement can cause debility and affect person's ability to carry out daily actions (eating, dressing, playing, personal hygiene etc.) and work [3]. Common treatments comprise nonsteroidal anti-inflammatory painkillers, intra-articular(joint) glucocorticoid steroids, oral glucocorticoids and also physiotherapy. Physically investigating the shoulder produces exact symptoms and also signs as a support for examiners in analysing the shoulder pain. There are different special tests for physically examining but making the selection of which special tests to use is complicative. Special tests are used to confirm a cautious diagnosis, supporting in the various diagnosis procedure. Clinical diagnosis makes the examiner to begin a prognostication and select proper interventions, finally getting the patient with best result. Orthopaedic special tests are performed after having complete examination of patient that includes patient characteristics, cause of injury, clinical examination, inspection, fleshy and bony tissue palpation, examination of active and passive range of motions, examination of passive joint movement, practical examinations and nervous related examination [4].

This article describes how regression algorithms like multi linear regression (MLR), decision tree regressor (DT), support vector regressor (SVR), gradient boosting regressor (GBT) and adaboost regressor (AdaR) can be used to predict the type of Shoulder disorders based on Orthopaedic special tests. There by it can support the examiner in making decision for further tests that depends on and make well-versed conclusions about their clinical diagnosis. The paper has conducted the programming analysis on the models based on the Python3.6 to predict type of shoulder disorder.

## 2 Shoulder Disorders

The shoulder is a multipart joint formed with humerus, scapula which are surrounded by ligaments, muscles and tendons. They assist the bones and attaches the clavicle, humerus and scapula. They also provides the glenoid cavity, acromion and coracoid processes. The two important joints of the shoulder are glenohumeral joint and acromioclavicular joint. Bursae provides flat sliding between bones, muscles, and tendons as shown in the Fig. 1 [1].

### 2.1 Rotator Cuff Disorders

The Rotator cuff muscles are group of four muscles that are located around the shoulder joint namely Supraspinatus, Infraspinatus, Subscapularis and Teres minor as shown in Fig. 2 [5]. Rotator cuff can be developed due to degeneration, swelling or trauma. Repair mechanisms fails and develops micro-tears which may further become macro-tears, and this leads to painful shoulder symptoms [6]. Treatment includes icing, rest, injections, anti-inflammatory medications, surgery, Physiotherapy.

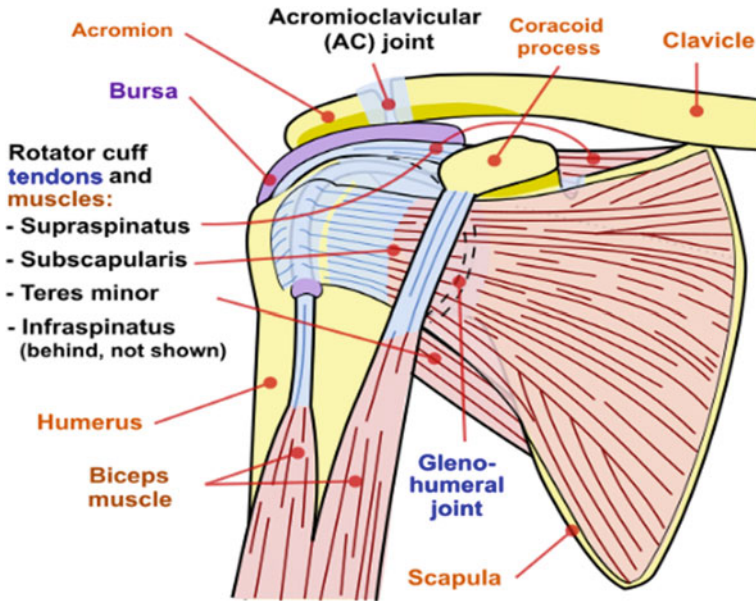


Fig. 1 Shoulder anatomy

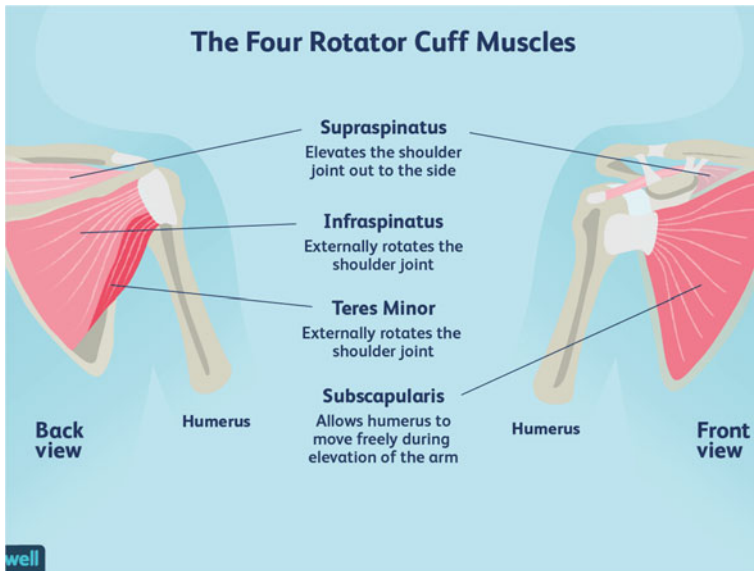


Fig. 2 Rotator cuff muscles

**Subscapularis tear.** Direct physical examination, including special tests like lift off test, belly press test, belly off test, passive lift off, bear-hug test, dropping sign tests can assist to identify tears of the subscapularis. This tear caused when arms are overextended and due to age-related degeneration.

**Infraspinatus tear.** Direct physical examination, including special tests like external rotation lag sign and Hornblower's sign test can be used to assist to identify tears of the Infraspinatus. The infraspinatus muscle is prone to trauma and to wear and tear from overuse and misuse.

**Supraspinatus tear.** Direct physical examination, including special tests like painful arc, drop arm, empty can, push-off and resistance, full can and external rotation lag sign can be used to assist to identify tears of the Supraspinatus. This tear occurs due to injuries and degeneration.

**Teres minor.** Direct physical examination, including special tests like hornblower's sign can be used to assist to identify tears of the Teres minor. This tear may occur due to outstretched arm or an attempt to lift heavy things.

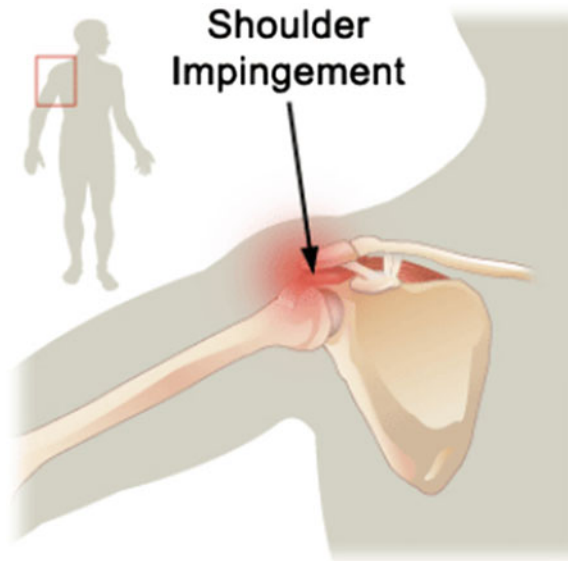
## 2.2 AC Joint Separation (ACJ)

ACJ is a separation between the upper part of the shoulder blade and the collarbone as shown in Fig. 3. AC joint separation occurs when the ligaments of this joint become damaged or torn. This causes a separation between the acromion and the collarbone [7]. It is caused due to directly falling onto the shoulder, being hit on the point of the shoulder blade or due to falling on an outstretched arm. Treatment includes icing, rest, anti-inflammatory medications, surgery. Special tests like painful arc, forced adduction with hanging, cross arm tests are used to assist to identify AC Joint Separation.

Fig. 3 AC Joint separation



**Fig. 4** Impingement syndrome



### **2.3 Impingement**

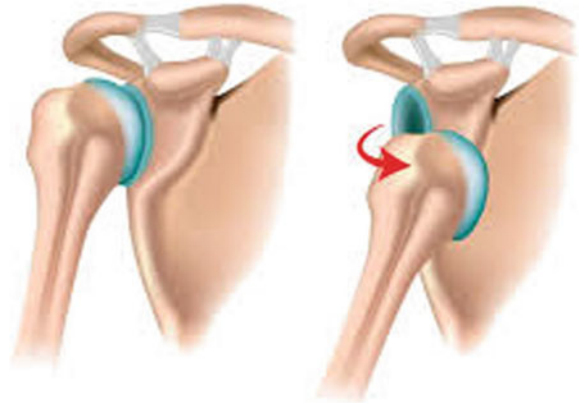
The impingement syndrome is a disorder where bursa and rotator cuff muscles or tendons are intermittently jammed and trampled during elevation movements of shoulder as shown in Fig. 4. This causes painful shoulder movements in the primary case but can turn to weakness which leads to rotator cuff disorder. Treatment includes icing, anti-inflammatory medications, and a systematized physical therapy program [8]. Special tests like neer, Hawkins Kennedy tests are used to assist in identifying Impingement syndrome.

### **2.4 Instability**

Shoulder instability arises when the structural parts that surround the shoulder joint do not work to maintain the ball within its socket, as a result the joint becomes too loose that it may slide partially out of place, this is called a shoulder dislocation as shown in Fig. 5. It commonly results from seizures, shock and falls, repetitive extreme external rotation with the humerus abducted and extended. Treatment includes closed reduction, surgery, shoulder brace, medication [9]. Special tests like Anterior Apprehension, Posterior apprehension, anterior posterior drawer, inferior instability test, sulcus, load & shift tests are used to assist in identifying Instability.



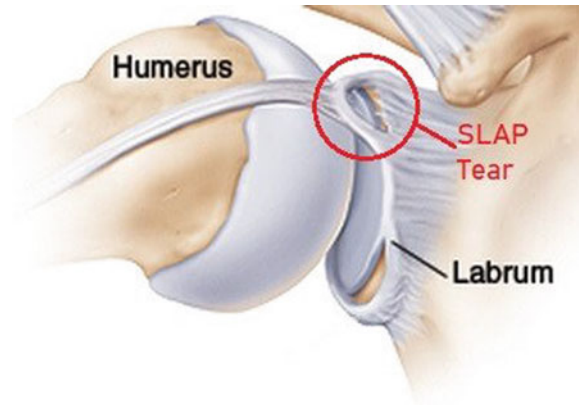
**Fig. 5** Instability



### 2.5 Labrum Tear

Labrum tears are of different types. It surrounds the shoulder socket and keeps the joint steady. Labrum tear means that it has torn at the top in both the front and back of shoulder where it attaches the biceps tendon as shown in Fig. 6. Labrum tear is caused as a result of trauma, structural abnormalities and repetitive motions. Labrum tear causes restlessness and limit the daily activities, and also to several sports activities [10] Treatment are Anti-inflammatories, rest, Physiotherapy [11]. Special tests like O brien, crank, SLAP prehension, clunk tests are used to predict Labrum tear.

**Fig. 6** Labrum tear



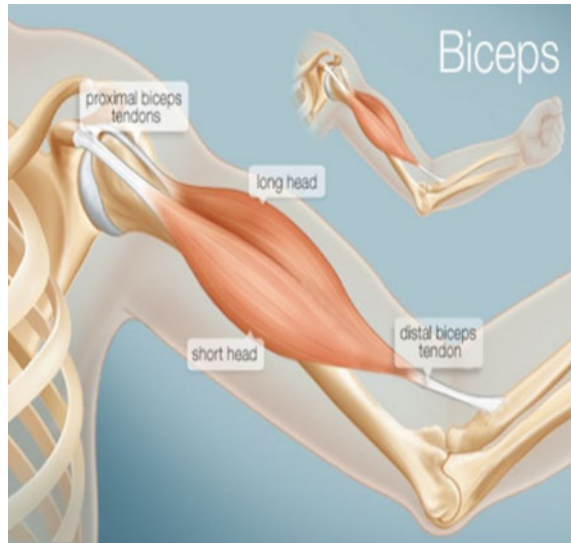
## 2.6 Biceps Tendon

The biceps are a muscle consists both short and long heads that work as a single muscle. It is attached with the arm bones by tough connective tissues called tendons. The tendons connect the biceps muscle to the shoulder joint in two places called the proximal biceps tendons That connects it to the bones of shoulder and distal biceps tendon that connects it to the radius bone at elbow [12] as shown in Fig. 7. Biceps tendon is caused as a result of sudden or serious load to the tendon. Treatment incudes cold packs, Nonsteroidal anti-inflammatory medications, rest, physiotherapy, corticosteroid injections and some cases includes surgery.

## 3 Regression Algorithms

Regression algorithms are supervised Machine Learning algorithms. These are used to find correlation between different variables. These algorithms are used to find the output values which depends upon the input values. These types of algorithms are used when the output value is real or continuous.

Fig. 7 Biceps tendon



### 3.1 *Multivariable Regression Algorithm*

This type of model explains the correlation between more than one independent variable and one dependent variable. A dependent variable is defined as a function of more than one independent variables with corresponding coefficients. It is defined as an equation containing dependent variable and more than one independent variables [13]. Multivariable Regression expression is expressed in Eq. 1.

$$Y_i = m_1x_1 + m_2x_2 + m_3x_3 + \dots + m_nx_n + b \quad (1)$$

where  $m_i$ 's ( $i = 1, 2, \dots, n$ ) referred as regression coefficients, that denotes the value at which the dependent variable ( $y$ ) varies when independent variables ( $x_{in}$ ) changes and  $b$  is constant.

### 3.2 *Decision Tree Regressor*

DTR perceives features of an object and trains a model in the form of tree structure predict the continuous output. It normally breakdowns the dataset into much smaller subsets and at the same time related decision tree is developed incrementally. Finally, tree is formed with nodes for making decision and also have leaf nodes.

### 3.3 *Gradient Boosting Regressor*

GBR is referred as integrated model having higher performance and better constancy which produces a strong model for prediction. It tries to minimise the loss function as expressed in Eq. 2. Friedman has proposed the GBR which has extended as boosting algorithm to solve various regression problems [14].

$$\text{Loss} = \text{MSE} = \sum (y_i - y_{i_p})^2 \quad (2)$$

where  $y_i$  =  $i$ th target variable,  $y_{i_p}$  =  $i$ th prediction variable, the loss function is  $L(y_i - y_p)$ .

### 3.4 *AdaBoost Regressor*

AdaBoost is one among the boosting algorithms. AdaBoost regressor is referred as meta estimator which first fits a regressor on the original dataset and then fits extra replicas of the regressor on the common dataset. The weights of instances are

adjusted depending on the error of the present prediction. Noisy values and outliers should be avoided while using ABR as it is sensitive to this data.

### 3.5 Support Vector Regressor

SVM was initially used to solve classification problems but it also used to resolve regression problems. SVR technique depends on kernel functions. Various kernel methods includes Sigmoid, Polynomial, Linear and Radial Basis. The user has to select the appropriate kernel function while building the model. Support vector regression uses epsilon loss function to answer the problems. It tries to fit error within a certain threshold.

## 4 Proposed Methodology

The data relating to shoulder disorder of patient was collected and pre-processed then trained with regression algorithms to predict the type of shoulder disorder. The following Fig. 8 shows the organization of Shoulder data.

### 4.1 Shoulder Dataset

Dataset of patients suffering from Shoulder disorders like rotator cuff disorders, acromioclavicular joint (ACJ), Instability, Impingement, labrum tear and Biceps tendon is collected. Shoulder dataset includes following variable list shown in the Table 1.

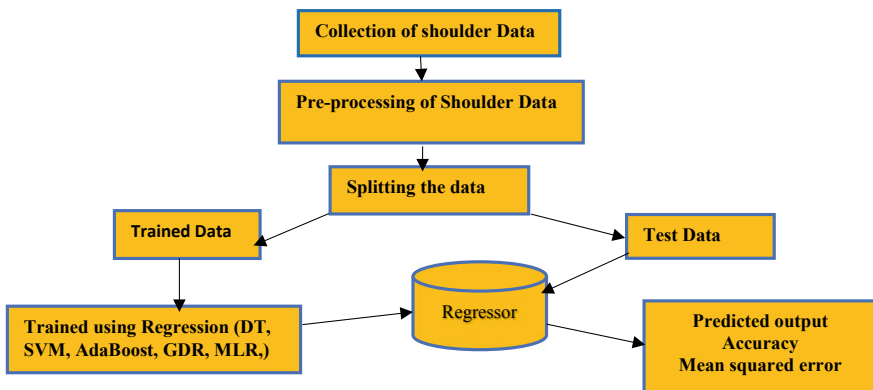


Fig. 8 Organization of data

**Table 1** Variable list of patients

Variable	Type	Code
Age	integer	Continuous values
Gender	Boolean	2 values (1 for male, 2 for female)
Smoke	Boolean	2 values (1 for +ve, 0 for -ve)
Shoulder	Boolean	2 values (1 for +ve, 0 for -ve)
Weakness	Boolean	2 values (1 for +ve, 0 for -ve)
Numbling	Boolean	2 values (1 for +ve, 0 for -ve)
Aching	Boolean	2 values (1 for +ve, 0 for -ve)
Sharp pain	Boolean	2 values (1 for +ve, 0 for -ve)
Night pain	Boolean	2 values (1 for +ve, 0 for -ve)
SPADI score	Integer	Continuous values
X-ray	Boolean	2 values (1 for +ve, 0 for -ve)
MRI	Boolean	2 values (1 for +ve, 0 for -ve)
Origin	Integer	9 codes (each for particular origin)
Painful-arc test	Boolean	2 values (1 for +ve, 0 for -ve)
Forced-adduction test	Boolean	2 values (1 for +ve, 0 for -ve)
Forced-adduction test in hanging	Boolean	2 values (1 for +ve, 0 for -ve)
Cross-arm test	Boolean	2 values (1 for +ve, 0 for -ve)
Neer-test	Boolean	2 values (1 for +ve, 0 for -ve)
Hawkin's-Kennedy test	Boolean	2 values (1 for +ve, 0 for -ve)
Anterior-Apprehension test	Boolean	2 values (1 for +ve, 0 for -ve)
Posterior-Apprehension test	Boolean	2 values (1 for +ve, 0 for -ve)
Anterior-Posterior drawer test	Boolean	2 values (1 for +ve, 0 for -ve)
Inferior-instability test	Boolean	2 values (1 for +ve, 0 for -ve)
Sulcus-test	Boolean	2 values (1 for +ve, 0 for -ve)
Load and shift test	Boolean	2 values (1 for +ve, 0 for -ve)
O'Brien-test	Boolean	2 values (1 for +ve, 0 for -ve)
Crank-test	Boolean	2 values (1 for +ve, 0 for -ve)
SLAP-prehension test	Boolean	2 values (1 for +ve, 0 for -ve)
Clunk-test	Boolean	2 values (1 for +ve, 0 for -ve)
Speed-test	Boolean	2 values (1 for +ve, 0 for -ve)
Yergasons test	Boolean	2 values (1 for +ve, 0 for -ve)
Lippmans test	Boolean	2 values (1 for +ve, 0 for -ve)
Drop-arm test	Boolean	2 values (1 for +ve, 0 for -ve)
Empty-can test	Boolean	2 values (1 for +ve, 0 for -ve)

(continued)

**Table 1** (continued)

Variable	Type	Code
Push-off and resistance test	Boolean	2 values (1 for +ve, 0 for -ve)
External-rotation lag sign	Boolean	2 values (1 for +ve, 0 for -ve)
Hornblower's-sign	Boolean	2 values (1 for +ve, 0 for -ve)
Lift-off	Boolean	2 values (1 for +ve, 0 for -ve)
Passive-lift-off	Boolean	2 values (1 for +ve, 0 for -ve)
Belly-off test	Boolean	2 values (1 for +ve, 0 for -ve)
Belly-press	Boolean	2 values (1 for +ve, 0 for -ve)
Bear-hug	Boolean	2 values (1 for +ve, 0 for -ve)
The dropping sign	Boolean	2 values (1 for +ve, 0 for -ve)

## 4.2 Pre-processing of Shoulder Dataset

Dataset is cleaned to avoid null values and missing values which are not supported by most of the machine learning algorithms.

## 4.3 Splitting of Shoulder Dataset

The whole dataset is divided into train and test data after selecting the independent and target variable. The 80% data was taken as training data and 20% of data was taken as testing data.

## 4.4 Regression

The training data is trained by using five machine learning algorithms i.e., Multi-variable Regression Algorithm, Decision Tree Regressor, Support Vector Regressor, Adaboost Regressor and Gradient boosting regressor

# 5 Experimental Results

The models gained learning from the training data set. In this learning, five regression algorithms were applied to the test data set to predict the type of shoulder disorder. The evaluation of machine learning models was based on the prediction accuracy, mean square error. The following figures shows the comparison between predicted value and actual value. In the following figures the type of shoulder disorders are

represented by numeric numbers as 1-Supraspinatus tear, 2-Infraspinatus tear, 3-Subscapularis tear, 4-Minor teres tear, 5-AC joint Separation, 6-Impingement, 7-Instability, 8-Labrum tear and 9-Biceps tendon.

**Multivariable Regression Algorithm.** The trained and test data is fitted using multivariable regression algorithm and type of shoulder disorder is predicted as shown in the Fig. 9. Depending upon the special tests the target variable was predicted, When Lift-off, Passive lift-off, Belly off sign, Belly press, Bear hug and the dropping sign tests were positive and other tests were negative, the type of shoulder was predicted as 3-Subscapularis tear. When Neer and Hawkins Kennedy tests were positive and other tests were negative the type of shoulder was predicted as 6-Impingement syndrome. This algorithm predicted all other shoulder disorders correctly depending upon the positive and negative of special tests.

**Decision Tree Regressor.** The trained and test data is fitted using Decision Tree Regressor and type of shoulder disorder is predicted as shown in the Fig. 10. When Speed, Yergason’s and Lippman’s tests were positive and other special tests were

Fig. 9 Prediction of type of shoulder using multivariable regression

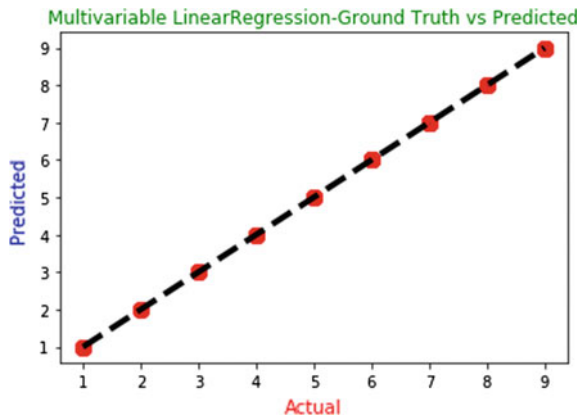
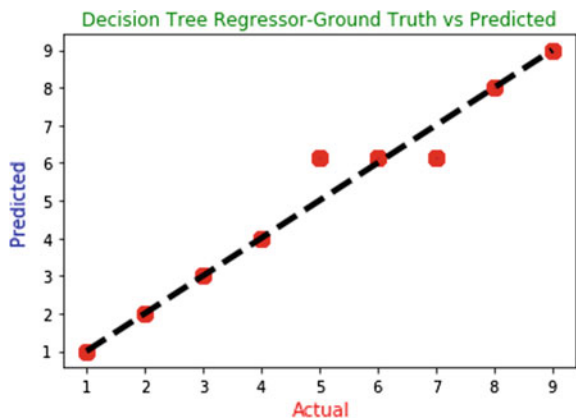
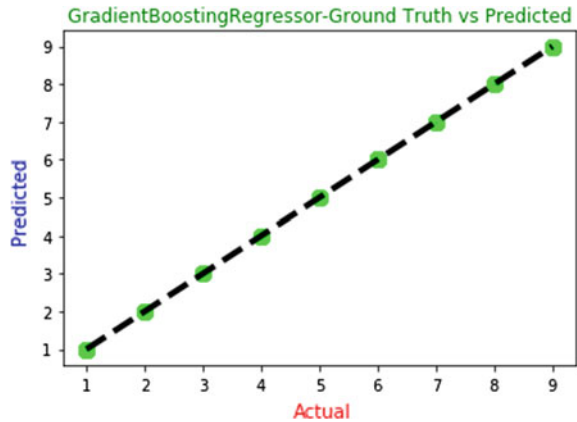


Fig. 10 Prediction of type of shoulder using decision tree regressor



**Fig. 11** Prediction of type of shoulder using gradient boosting regressor



negative, the type of shoulder predicted as 9-Biceps tendon. When Hornblowers sign and External-rotation lag sign are positive and other tests are negative, the type of shoulder predicted as 2-Infraspinatus tear. The disorders Instability and AC joint Separation shoulder disorders were not predicted accurately but all other disorders were predicted correctly.

**Gradient Boosting Regressor.** The trained and test data is fitted using Gradient boosting regressor and type of shoulder disorder is predicted as shown in the Fig. 11. When Painful arc, Drop-arm, Empty can, Push-off& Resistance, Full can and External rotation lag sign were positive and other special tests were negative, the type of shoulder predicted as 1-Supraspinatus tear. When Hornblower’s sign is positive and other special tests were negative the type of shoulder predicted as 4-Minor teres tear. This algorithm predicted all other shoulder disorders correctly depending upon the positive and negative of special tests.

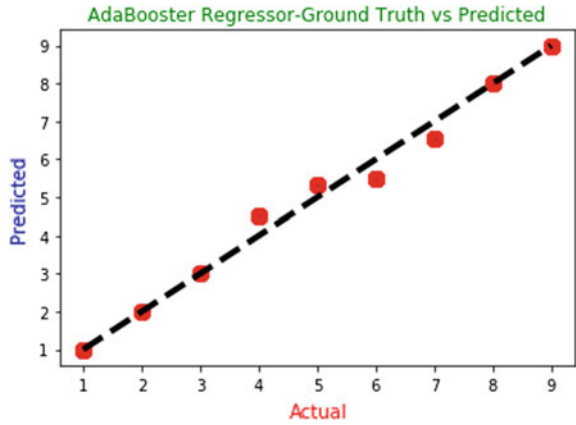
**Adaboost Regressor.** The trained and test data is fitted using AdaBoost Regressor and type of shoulder disorder is predicted as shown in the Fig. 12. When Painful arc, Forced Adduction, cross arm tests are positive and other special tests were negative, the type of shoulder predicted as 5-AC joint Separation. The disorders Minor teres tear, Instability and Impingement shoulder disorders were not predicted accurately but all other disorders were predicted correctly.

**Support Vector Regressor.** The trained and test data is fitted using Support Vector Regressor and type of shoulder disorder is predicted as shown in the Fig. 13. When Anterior-Apprehension test, Posterior-Apprehension test, Anterior-Posterior drawer test, Inferior-instability test, Sulcus and Load and Shift are positive and other special tests were negative, the type of shoulder predicted as 7-Instability. This algorithm predicted all other shoulder disorders correctly depending upon the positive and negative of special tests.

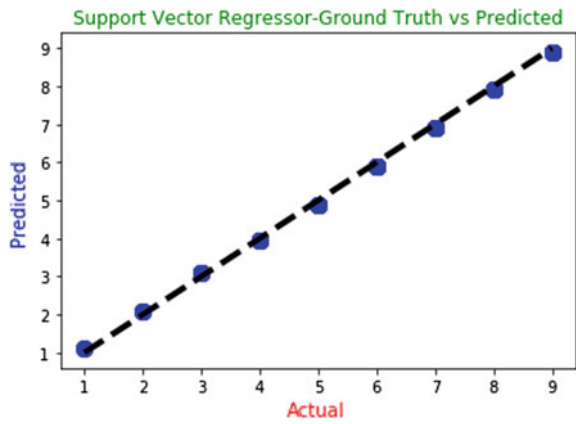
The following Table 2 and Fig. 14 shows the comparison of the results obtained from five regression algorithms. For automated diagnosis of shoulder pain, Multi-variable regression, Gradient boosting regressor, and Support vector regressor had the highest accuracy than decision tree regressor and adaboost regressor.



**Fig. 12** Prediction of type of shoulder using adaboost regressor



**Fig. 13** Prediction of type of shoulder using support vector regressor



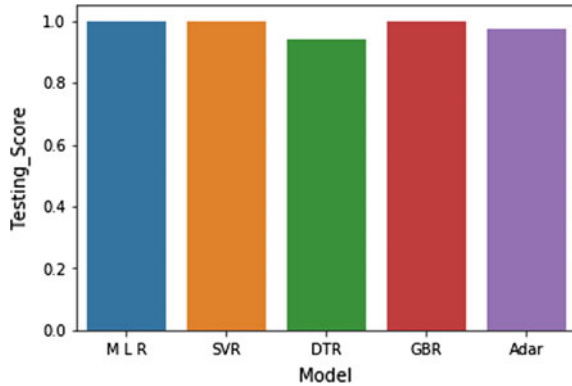
**Table 2** Prediction performance of models

Model	Training score	Testing score	MSE	Accuracy (%)
DTR	0.960218	0.949108	0.29	95
AdaR	0.980912	0.980816	0.11	98
SVR	0.998420	0.998283	0.01	99
GBR	0.9999999	0.9999999	0.00	99
MLR	1.000000	1.000000	0.00	100

## 6 Conclusion

Shoulder pain is one of the important and commonly seen musculoskeletal problem. The lack of consensus on diagnostic measures and clinical assessment may complicate treatment choices. This paper helps to confirm a cautious diagnosis,

**Fig. 14** Accuracy result between regressor algorithms



supporting in the various diagnosis procedure, differentiating among and between several potentially compulsive tissues. This could help the practitioner to take further interventions. Regression algorithms assist in finding the hidden information and knowledge in data that predict diagnosis and treatment of diseases in Shoulder based on special tests. This paper implemented five machine learning techniques by using the shoulder dataset to predict type of shoulder disorder. Comparative analysis using python with regression algorithms we found that Support Vector Regressor, Multivariable Regression and Gradient Boosting Regressor had the highest accuracy to predict the type of shoulder pain.

## References

1. <https://en.wikipedia.org/wiki/Shoulder>
2. T.D. Tennent, R.B William, F.M. John, A review of the special tests associated with shoulder examination: part I: the Rotator Cuff tests. *Am. J. Sports Med.* **31**(1), 154–160 (2003)
3. M. Caroline et al., Shoulder pain: diagnosis and management in primary care. *BMJ (Clinical research ed.)* **331**(7525), 1124–8 (2005). <https://doi.org/10.1136/bmj.331.7525.1124>
4. N.E. Biederwolf, A proposed evidence-based shoulder special testing examination algorithm: clinical utility based on a systematic review of the literature. *Int. J. Sports Phys. Ther.* **8**(4), 427 (2013)
5. <https://www.verywellhealth.com/the-rotator-cuff-2696385>
6. C.H. Linaker, K. Walker-Bone, Shoulder disorders and occupation. *Best practice & research. Clinical rheumatology* **29**(3), 405–423 (2015). <https://doi.org/10.1016/j.berh.2015.04.001>
7. <https://www.healthlinkbc.ca/health-topics/zm6359>
8. <https://physioworks.com.au/injuries-conditions-1/rotator-cuff-impingement>
9. <https://www.sportsinjuryclinic.net/sport-injuries/shoulder-pain/acute-shoulder-injuries/dislocated-shoulder>
10. R. Vander Kraats, A. Doss, Glenoid labral tear follows up case series on ultrasound guided autologous platelet rich plasma in conjunction with a progressive rehabilitation program. *F1000Res* **1**, 68 (2012)
11. <https://www.shoulder-pain-explained.com/SLAP-tear.html>
12. <https://www.webmd.com/fitness-exercise/picture-of-the-biceps#1>

13. C. Combes, F. Kadri, S. Chaabane, Predicting hospital length of stay using regression models: application to emergency department. (2014)
14. Xingyan Li, Weidong Li, Xu Yan, Human Age prediction based on DNA methylation using a gradient boosting regressor. *Genes* 9(9), 424 (2018)

# Diagnosis of Urological Diseases Using Deep ROI



R. Venkata Raviteja, M. Abhilaasha, and B. Prakasha Rao

**Abstract** Artificial intelligence (AI)—is the capability of device to accomplish cognitive chores to realize a specific aim to support provided data is transformed and reformed our health care schemes. The existing disposal of computational authority is increased, extremely advanced for recognition of patterns and software for image processing is advanced for operation. The hustles for improvement leads to advent of computer based schemes which are skilled to accomplish difficult chores in robotics and images related to medicine. The Urological disease is a situation connected for purifying and transport of urine out of the body. These diseases occur in humans like male, female and children of all age. This will effect at particular parts of the body. in females, it involves in urinary tract. In men's, it effects on urinary tract/reproductive organs. The urinary tract is the body's drainage structure for eliminating urine. Urine is collection of trashes and liquid. The urinary tract contains bladder, kidney and ureters. In generally to urinate, urinary tract wants to work in exact direction. The urological diseases can be specified by observing the prostate problems, bladder control problems, urinary tract infections, kidney stones et al. mostly urological disorders are short in time or may be long time. In this paper, a novel approach is exhausted with feature extraction and deep Region of Interest (ROI) for diagnosis of urological diseases is proposed. They claim that two segments are used for extracting. The segments are as follows: (i) A deep learning architecture of a pre-defined training data used for feature extraction (ii) to classify, machine learning. In order to estimate our suggested method, we calculated the Intersection over Union (IoU) for our extracted ROI with accurateness, Receiver Operating Characteristic (ROC) curves, and Equal Error Rate (EER) for conformation chore. The investigation explains about extraction module of ROI can expressively find the applicable ROI's and the corroboration results were critically extracted.

---

R. Venkata Raviteja (✉) · M. Abhilaasha · B. Prakasha Rao  
Guntur Medical College, Guntur, India  
e-mail: [raviteja091@gmail.com](mailto:raviteja091@gmail.com)

M. Abhilaasha  
e-mail: [abhilaashart@gmail.com](mailto:abhilaashart@gmail.com)

B. Prakasha Rao  
e-mail: [raopbusam@gmail.com](mailto:raopbusam@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_45](https://doi.org/10.1007/978-3-030-46939-9_45)

**Keywords** Urology · Image processing · Region of interest · Deep learning

## 1 Introduction

The Artificial Intelligence (AI) has vast explosion in sharing and claiming in health care. The medical data has huge volume of records are available for the advent of computer medical records are used to decrease difficulty in medical testing. AI is planned to reform our consumption of medical data. Health care AI stock is growing, totally about \$2.14 billion over 323 deals since 2012–2017 and has improved each year, counting via 31% in 2016, rendering to a 2017 CB Perceptions research report.

For diagnosis [1] management [2] and outcome prediction [2, 3] of urological diseases are gradually applied to AI in the earlier decade. In current analysis we concisely familiarize AI technologies focuses on numerous applications of assorted AI algorithms in the supervision of urological diseases. Compare AI methodologies for predicting normal and abnormal images of Computed Tomography (CT) scan images by Region of interest.

### 1.1 About Urological Diseases

Urological diseases, disorders, and situation involves persons [4] of all age, effect in major health care expenditures, which might lead to significant infirmity and impair worth of life. Non-cancerous (benign) urological health effects contain urinary tract infections, kidney stones, urinary incontinence, and benign prostatic hyperplasia [5] (an enlarged prostate). Interstitial cystitis/bladder pain syndrome (IC/BPS), a devastating and hurting state, researchers estimates approximate 3.3 million women, and 1.6 million men are affected with chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS) consisting of urological symptoms like pain with bladder filling [6]. Based upon national public health surveys conducted over numerous years, it is approximately 1 in 10 U.S. adults (18 years of age and older) hurt from daily urinary incontinence; mainly it affected by women. A lot of people undergo in silence due to shame and lack of information about accessible treatment options.

In health care applications AI remains mystery. The Oxford Dictionary define AI as “the theory and development of computer systems able to perform tasks that normally require human intelligence.” In health care, AI is not explicitly defining but normally it uses machine learning algorithms for entire tasks like drug discovery, virtually assisting patients, and automating complex medical tasks such as analysis of diagnostic tests.

Urological practices have been applied by AI three general areas like pathologic diagnosis, radiologic, telemedicine, Patient monitoring and automating respective task. It has been tested in other two areas like precision medicine, data analytics and surgical training, quietly improvement (Table 1).

**Table 1** Different scenarios in medical using AI

Areas	Scenarios	AI related technologic	Public companies	Start-ups
Imaging department	Disease diagnosis	Computer vision, deep learning, machine learning	Google, tencent	CureMetrix, zebra
	Medical equipment upgrade	Computer vision	GE, siemens	TSMS, 3DHISTECH
	Medical imaging information mining	Data mining, machine learning	IBM, NVIDIA	Enlitic, 12sigma
	Digital pathology	Computer vision, deep learning	GE, google	Cirdan, deep lens
Operating department	Surgical robot	Robotics, computer vision	Zimmer biomet, intuitive surgical	Pathfinder technologies, medrobotics
	Surgical planning and navigation	Computer vision, AR	Stryker, medtronic	Onkos surgical, SiMMo3D
	Tumor radiotherapy	Deep learning, computer vision	Siemens, varian	Panacea, siris medical

**Medical Image Process Energizes AI:** The tender of computer vision technology is the basic form of Medical Image Process Machine learning and deep learning are the intellectual analysis of AI technologies [7] which are processes by computer vision technologies like image registration fusion, and doctors can get assistant, from medical image labelling, disease diagnosis, and surgery.

**Analysis of data types integrated with other medical images:** The medical image date has collection of patients physical signs, history of reports, genetic information and non-image data for the training of AI algorithm. This helps to analyse machine data which has higher dimensions and except most vital features. By Reconnoitring implicit connections of diseases can help the doctors to diagnose diseases more exactly.

**For Diagnoses the Disease Machine Learning not exceeds Deep learning:** Predictable CAD method is mostly based on machine learning [8] or systems are criticized for running unthinkingly or helpful. Moreover, deep learning algorithms are proven more capable for dealing with medical data widely, extracting valuable data and outputting main states of disease. Which releases doctors from time wasting clinical work.

## ***1.2 Scenarios Usages of Medical Images in AI***

The main advantage of AI will deal with medical image data as image processing instinctively use CNN and RNN adaptively. In clinics and other medical societies like liberated image centres and physical check-up centres, medical image process are one of the useful application of AI [9]. The Table 1 shows exclude scenarios of medical image data like medical text processing it neglects bioelectricity, where image processing methods are intensely diverse from normal computer vision methods.

## ***1.3 Restrictions and Tendencies of AI in Medical Image***

In medical image AI technology is welcomed in market to reduce doctors time and inefficiencies but it limits the factors for practical application of AI medical imaging. For illustration different hospitals, research institutions, grim to assemble and exploit medical image data effectually.

AI remains to push medical image technology rapidly to develop and deploy. To diagnoses disease, AI is applied in cellular/molecular level image process, interventional image, assisting non-surgical diagnosis and treatment. It regulates the industry association and cooperate organizations of medical images to inaugurate algorithm for evaluate standards in the industry.

## **2 Image Processing**

Medical image is a process of analysing inner parts of the body through creating visual demonstrations as well as pictorial demonstration of organs or tissue functions. Medical image will reveal the inner parts of the body which are hidden by bones and skin. These are used to diagnose the disease for treatment it also creates a database for anatomy or physiology to identify irregularities. Though by image we can remove organs and tissues for performance reasons in medical records such type of procedure is used by pathology instead of medical imaging.

Biological imaging and Radiology imaging are the two types of medical images which are widely used in X-ray radiography, medical photography, endoscopy, tactile imaging, ultrasonography or ultrasound, magnetic echo imaging, thermography and nuclear medicine functional imaging techniques as positron emission tomography (PET) and single-photon emission computed tomography (SPECT).

Measurement and recording methods aren't designed for the supply images like electroencephalography (EEG), magnetoencephalography (MEG), electrocardiography (ECG) et al., symbolized other skills produces data vulnerable to represent the

parameter graph vs time or maps. Which contains detailed information like dimensions of location during partial evaluation these skills are considered for the medical images in alternative discipline.

In worldwide 5 billion medical images are studied at 2010 [10]. In 2006, at united states medical images expose radiation which has 50% of total realising radiation [11]. The manufacturing equipment for medical image uses technology from the semiconductors, CMOS integrated circuit chips, power semiconductors, sensors like image sensors (mainly CMOS sensors) and bio sensors. Processors like microcontrollers, media processors, digital signal processors, microprocessors and system-on-chip devices. At 2015, 46 million units of medical image chips amounts \$1.1 billion on annual shipment [11].

Medical image habitually observed to design a set of methods to produce images non-invasively for the inner parts of body. In medical image we can see the solution for the restricted sense through mathematical inverse problems which means the living tissue (cause) and observed signal (effect). In Ultrasonography the pressure and echo of ultrasonic waves goes into the tissues to display inner structure. As in projectional radiography uses X-ray radiation which observed at dissimilar rates and dissimilar tissue forms like muscle, fat and bone.

The word “non-invasive” is denoted as procedure where instruments are not send into patient’s body. These methods are mostly used in image.

### 3 Edge Detection

For images detection of edges [12] is the primary and secondary order of derivate for an image. For illustration, the edges can be detected by looking for absolute values of the first derivative or crossing the zero value for the secondary derivative of the image [11]. Conventionally, edge detected based on the early predefined algorithms like Prewitt, Laplacian of Gaussian operator, Sobel and canny. But in theory these operators are belonged to the high pass filter which are not suitable for noisy images. In the practical view of point detection of edges are categorised into 2 forms: searched based and zero crossing based. The searched base method detects the edges and calculate the edges strength, like gradient’s magnitude, depth and local maxima to search in a way which matches by the gradient direction of edge outline. The zero crossing is a second calculation expression computed from images to find edges like laplacian or non-linear expression [13]. In the theoretical point of view recognition of edges are classified into contextual and non-contextual methodologies. The non-contextual will work alone without knowing any prior information about the edges and sections. This are flexible and not limited to the particular image. Through, processing is based on local areas and also focuses on the neighbouring pixel areas. The contextual method directed by priori information about edges/sections. This will perform exactly in a specific context. It clears independent detectors are apt for overall purpose claims. Through contextual detector adapts the specific application which include images by similar sections or objects.



- Smoothing: Suppress the maximum amount noise as possible without destroying the truth edges.
- Enhancement: Application of a filter to reinforce the standard of the sides within the image.
- Detection: It determines edges of a pixel to remove the noise to retain image.
- Localization: It determines to exact the edges of an object which are thine and linking are required.

### 3.1 Threshold

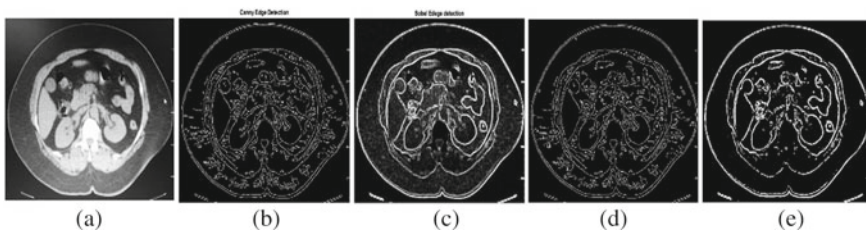
The segmentation [14] is more robust, the system will choose the knowledge about the objects, environment, applications are selected threshold automatically:

- Objects in depth features
- Objects Dimensions
- Sections of images occupied by the objects
- Number of various sort of objects appearing in image.

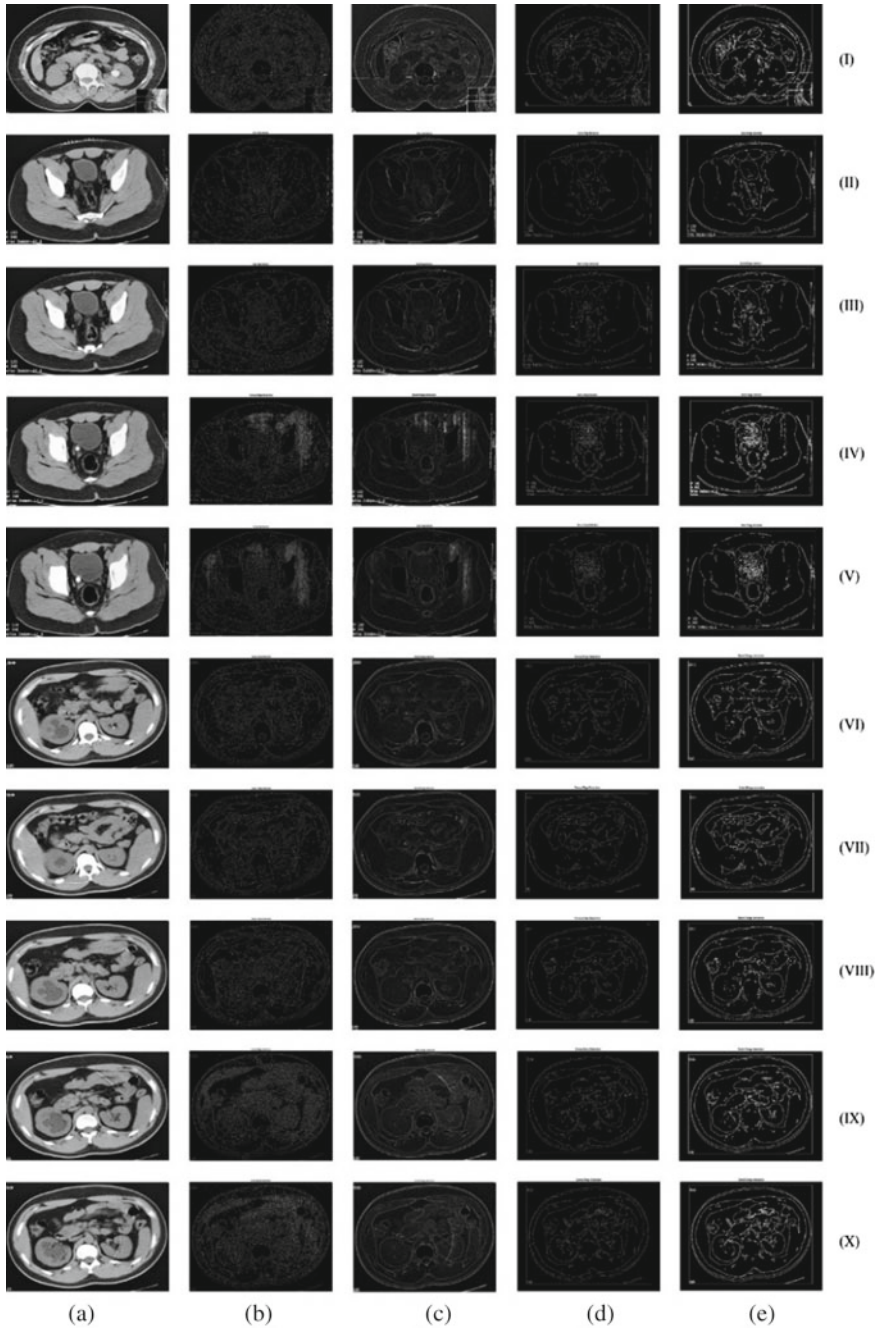
## 4 Experimental Analysis

Firstly, taken the urology Computed Tomography (CT) scan image for analysis to that image edge detection methods are applied on original image as shown in Fig. 1a. Edge detection methods of canny and sobel methods are applied as illustrated in Fig. 1b, c. for edge detection output Region of Interest(ROI) is applied individually for canny and sobel as shown in Fig. 1d, e.

For analysis around 10–15 images has been taken and applied the edge detection techniques and ROI for all images as shown in Fig. 2. To find the normal and abnormality of the image here taken the threshold values.



**Fig. 1** Normal image **a** input image **b** detection of edges using canny **c** detection of edges using sobel **d** ROI of canny **e** ROI of sobel



**Fig. 2** Abnormal images **a** input image **b** detection of edges using canny **c** detection of edges using sobel **d** ROI of canny **e** ROI of sobel

**Table 2** Threshold values for normal image

S. No.	Normal edge detection		ROI for edge detection	
	Canny	Sobel	Canny	Sobel
1	124	124	128	139

**Table 3** Threshold values for abnormal images

S. No.	Normal edge detection		ROI for edge detection	
	Canny	Sobel	Canny	Sobel
1	128	128	129	142
2	128	128	128	133
3	128	128	128	133
4	128	128	127	134
5	128	128	128	134
6	103	103	128	136
7	103	103	128	135
8	103	103	128	136
9	103	103	128	139
10	103	103	128	137

The threshold values of canny and sobel detection for edge is considered for our evaluation because of canny can give wide range of edges for a given image and Sobel performs a 2-Dimensional spatial gradient size for the image. In health care images tumour has the boundaries which is defined in image for measuring the sizes for that purpose threshold values of ROI are evaluated.

By using MATLAB, we evaluated the edge detection methods, ROI and considered threshold values. The threshold values of normal and abnormal images are differentiated for urology CT scan images through Canny, Sobel edge detection techniques. We applied ROI for canny and sobel edge detection for input images are as illustrated in Fig. 1, and threshold values of those images are shown in Table 2 for normal images and Table 3 for abnormal images.

## 5 Conclusion

The “urological diseases” is a situation connected for purifying and transports urine out of the body. This disease occurs in humans like Male, female and children of all age. This will effect at particular parts in the body. In females, it effects at urinary tract. In men’s it effects on urinary tract/reproductive organs. The urinary tract is a drainage structure for eliminating urine form the body. Taken the urology Computed Tomography (CT) scan image for analysis to that image edge detection methods

are applied. For analysis around 10–15 images has been taken and applied the edge detection techniques, ROI for all images. The normal and abnormal images are differentiated based on the threshold values of ROI images. These threshold value comparisons are going to use, and the next big era in AI and the technology that will bring immense ROI for our further research works.

## References

1. Y. Kanagasigam, D. Xiao, J. Vignarajan, Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. e182665 (2018)
2. S.J. Drouin, D.R. Yates, V. Hupertan, O. Cussenot, M. Roupret, A systematic review of the tools available for predicting survival and managing patients with urothelial carcinomas of the bladder and of the upper tract in a curative setting. *World J. Urol.* **31**, 109–116 (2013)
3. A.J. Hung, J. Chen, I.S. Gill, Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg* **153**, 770–771 (2018)
4. Urology Profile. Canadian Medical Association (2019), <https://www.cma.ca/sites/default/files/2019-01/urology-e.pdf>
5. X. Zhu, A.J. Klijn, L.M.O. de Kort, urological, sexual and quality of life evaluation of adult patients with extrophy-epispadias complex: long-term results from a dutch cohort (2019), <https://doi.org/10.1016/j.urology.2019.10.011>
6. S. Setia, C. feng, C. Coogan, S. Vourganti, M. Abern, in *Urology Residents Experience with Simulation: Initial Evaluation of MRI/US Fusion Biopsy Workshop*, vol. 134. pp. 51–55 (2019), <https://doi.org/10.1016/j.urology.2019.09.004>
7. S. Russell, J. Bohannon, Artificial intelligence. fears of an AI pioneer. *Sci.* **349**, 252 (2015). <https://doi.org/10.1126/science.349.6245.252>
8. F. Pesapane, M. Codari., F. Sardanelli, Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine (2018), <https://doi.org/10.1186/s41747-018-0061-6>
9. J.R. England, P.M. Cheng, Artificial intelligence for medical image analysis. *Am. J. Roentgenology.* **212**, 513–519 (2019)
10. C.A. Roobottom, G. Mitchell, G. Morgan-Hughes, Radiation-reduction strategies in cardiac computed tomographic angiography. *Clin. Radiol.* **65**(11), 859–67 (2010). <https://doi.org/10.1016/j.crad.2010.04.021.pmid20933639>
11. Medical Radiation Exposure of the U.S. population Greatly Increased Since the Early 1980's. National Council on Radiation Protection & Measurements. (2009)
12. G. Nagalakshmi, S. Jyothi, Edge detection methods for image segmentation. *Int. J. Comput. Sci. Mathe. Eng.* **1**(6) (2014). ISSN-2349–8439
13. Medical imaging chip global unit volume to soar over the next five years. *Silicon Semicond.* (2016)
14. P. Prathusha, S. Jyothi, A novel edge detection algorithm for fast and efficient image segmentation, in *Data Engineering and Intelligent Computing, Advances in Intelligent Systems and Computing.* vol. 542, pp. 283–291 (2017)

# Pharmacokinetic and Pharmacodynamic Studies on Celecoxib Loaded Nanosponges Gel for Topical Delivery



Y. Sarah Sujitha and Y. Indira Muzib

**Abstract** The present work objective is to formulate topical gel loaded with celecoxib nanosponges to enhance solubility and to reduce the side effects related to oral administration. Nanosponges loaded with celecoxib drug were formulated by using the emulsion solvent diffusion method. Nanosponges were subjected to various physicochemical parameters. The optimized formula of celecoxib nanosponges was loaded into gel base. The *ex vivo* drug release was studied by using modified Franz diffusion. Further they were subjected to pharmacokinetic and pharmacodynamic activity. The developed optimized celecoxib loaded nanosponges were analyzed by using SEM, FT-IR, DSC studies which shows that the prepared nanosponges reveals no chemical incompatibility between the polymer and drug. The size of the particle, zeta potential, and entrapment efficiency of the optimized nanosponges were found to be 240.9 nm and  $-1.8$  mV and 98% respectively. Optimized formulation X1 of gel appears homogenous, showing pH  $7 \pm 0.2$ , viscosity 8640cps, drug content 96.4% and *ex vivo* skin permeation studies found to be 95% for 4 h. Pharmacokinetic and pharmacodynamic results show that the celecoxib loaded nanosponges gel is more effective compared to marketed formulation. Celecoxib loaded nanosponges gel shows significant activity ( $p < 0.05$ ) compared to the Marketed formulation (voveran Emulgel).

**Keywords** Nanosponges · Nanosponges gel · Celecoxib · Pharmacokinetic and pharmacodynamic activity

## 1 Introduction

Nanosponges are tiny sponges with a size of about 250 nm-1  $\mu$ m have the capacity of loading or entrapping a wide range of drugs. To incorporate the drug into nanosponges the drug should have the molecular weight below 400 Daltons. Nanosponges can be

---

Y. Sarah Sujitha (✉) · Y. Indira Muzib  
Institute of Pharmaceutical Technology, Sri Padmavathi Mahila Visvavidyalayam (Women's University), Tirupathi, India  
e-mail: [suji.sarah@gmail.com](mailto:suji.sarah@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_46](https://doi.org/10.1007/978-3-030-46939-9_46)

525

easily incorporated into topical dosage form and effectively releases the drug with an increased safety and stability in an effective way and bypasses the first pass metabolism thus reduces the gastrointestinal side effects. These nanosponges due to its small particle size they can freely circulate in the blood until they stick onto a specific target site and release the drug [1, 2].

The commonly used drugs for the treatment of pain and inflammation caused by rheumatoid arthritis are NSAIDs. Celecoxib is a non steroidal anti-inflammatory drug (NSAID) 4-5 -(4-methyl phenyl) -3 (trifluoromethyl)-1H-pyrazole-1-yl benzene sulfonamides and it was the first synthesized non-steroidal anti-inflammatory drug. Celecoxib is a poorly water-soluble, low molecular weight, high lipid solubility drug with high permeation capacity [3]. Celecoxib is generally used for the treatment of osteoarthritis, acute pain and rheumatoid arthritis with oral route. But the long-term administration of this celecoxib leads to gastric side effects. To overcome these problems nanosponges was developed in which celecoxib was loaded and given topically. The present study is aimed to increase the solubility of poorly aqueous soluble celecoxib drug and to bypass the first pass metabolism by which it reduces gastric irritation and formulating it as nanosponges and converting it into the topical gel. Hence the objective of the present study is aimed to formulate, characterize, *in vitro* and *in vivo* evaluation of nanosponges loaded topical gel [4].

## 2 Materials

Celecoxib received as gift sample from Kekule Pharma LTD, Hyderabad, India. Hydroxy Propyl Methyl Cellulose (HPMC), Ethyl Cellulose (EC), Carbopol, Poly vinyl alcohol were purchased from Hi-Media laboratories Mumbai. All the solvents and reagents used were of analytical grade.

## 3 Methods

### 3.1 Formulation of Celecoxib Loaded Nanosponges

Nanosponges loaded with celecoxib loaded were prepared by emulsion solvent diffusion method using different concentrations of ethyl cellulose (shown in Table 1) [5]. The drug and polymer are added to organic phase (di chloromethane) and the external aqueous phase containing the water and PVA. Slowly the organic phase is added to the external phase with a constant stirring on magnetic stirrer (Remi equipments, Mumbai) at 1200 RPM for 30 min. Then formed nanosponges were filtered and dried using freeze dryer (Lyophilizer). Celecoxib loaded nanosponges optimization was given in Table 1.

**Table 1** Formulation and optimization of celecoxib loaded nanosponges

S. No.	Formulation code	Drug: EC mg)	PVA (mg)	Water (ml)
1	F1	1:1	100	150
2	F2	1:2	100	150
3	F3	1:3	100	150
4	F4	1 :4	100	150
5	F5	1:5	100	150
6	F6	1:4	100	200
7	F7	1:4	100	300
8	F8	1:4	150	150
9	F9	1:4	200	150

### 3.2 Formulation of Celecoxib Loaded Nanosponges Gel

Celecoxib loaded nanosponges gel was prepared by using combination of polymers HPMC and carbopol 934 [6, 7], composition was given in Table 2. Different concentrations of polymers were used. The celecoxib nanosponges gel was prepared initially by soaking carbopol 934, HPMC and optimized nanosponges formulation F4 overnight in required quantity of water (30 ml), and stirred by using mechanical stirrer at 1000 RPM for half an hour at temperature 37°C to get a homogenous dispersion, then methylparaben, propylparaben (preservatives), glycerin, tri ethanolamine (to neutralize the pH) was added to the gel and stirring was continued for 30 min. Optimization of celecoxib loaded nanosponges gel was given in the Table 2.

**Table 2** Formulation and optimization of topical gels using polymers (carbopol934 and HPMC)

S. No.	Formulation code	Composition	Ratios of polymers
1	X1	Carbopol: HPMC	1:1
2	X2	Carbopol: HPMC	1:2
3	X3	Carbopol: HPMC	1:3
4	X4	Carbopol: HPMC	1:4
5	X5	Carbopol: HPMC	2:1
6	X6	Carbopol: HPMC (pure drug)	1:1

## **4 Evaluation of Celecoxib Loaded Nanosponges**

### **4.1 SEM Analysis**

Scanning electron microscopy analysis was used for the study of surface morphology and shape of the celecoxib loaded nanosponges [8] (JOEL-JSM-IT500). The captured images were elaborated by using processing program and shape and surface morphology of the particles were observed.

### **4.2 Particle Size and Zeta Potential**

The particle size and zeta potential of the celecoxib nanosponges were analyzed by using zetasizer (Horiba) [9] to know the size range of particles and stability of the dispersion. Samples were measured in triplicates and the average was taken.

### **4.3 Differential Scanning Calorimeter**

DSC studies were performed to find out the chemical compatibility of drug and polymer (ethyl cellulose and celecoxib) [9] (Mettler Toledo, Switzerland). Drug and formulation were heated gradually at a rate of 10°C/min under constant heating in a temperature range of 30–200 °C.

### **4.4 Fourier Transform Infra Red Spectroscopy (FT-IR)**

FT-IR analysis was performed for the verification of any chemical interference between the drug and polymer (FTS300 Spectrophotometer (DGLab), canton, MA). Formed pellet was scanned in the range of 400–4000 cm<sup>-1</sup> [4].

### **4.5 Powder X-Ray Diffraction Studies**

The crystalline nature of the pure celecoxib and celecoxib nanosponges powder was evaluated by using powder X-ray diffractometer (Rigaku mini plus) [10, 11]. The rate of scanning employed was 7° min over 10–40 diffraction angle of 2θ range.



#### **4.6 Entrapment Efficiency**

Entrapment efficiency of celecoxib loaded nanosponges was determined by taking 10 mg of celecoxib loaded nanosponges in 10 ml phosphate buffer of pH 7.4 [12, 13] and the solution was kept aside for 24 h and after 24 h the solution was sonicated for five minutes using probe sonicator and stirred for 30 min by using a magnetic stirrer (Remi equipments, Mumbai). The solution was filtered and absorbance of the above solution was observed by using a UV spectrophotometer at 251 nm (shimadzu, Japan).

#### **4.7 In Vitro Drug Release Studies of Celecoxib Loaded Nanosponges**

The *in vitro* drug release studies were performed using the dialysis membrane method containing a pore size of 2.4 nm and molecular weight of 12,000–14,000 Daltons was used [14]. Phosphate buffer pH 7.4 was used to fill the dialysis membrane, 100 mg equivalent weight of celecoxib loaded nanosponges. The beaker was maintained at 37 °C temperature and rotated at 100 RPM. At regular time intervals samples were collected as 0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 12, 24 h and equal amount of the freshly prepared buffer sample was replaced into beaker maintain sink condition and the samples were measured by UV Spectrophotometer (shimadzu, Japan) at the wave length 251 nm.

### **5 Physicochemical Evaluation of Celecoxib Loaded Nanosponges Gel**

#### **5.1 PH of the Gel**

The pH of the celecoxib loaded nanosponges [14] gel was determined by using digital pH meter. The pH of the gel was checked at 1st and the 15th and 30th day after the preparation of gel to find any changes with reference to the time.

#### **5.2 Viscosity**

Five grams of gel was taken in a beaker and the viscosity was checked by using a brook field viscometer (Brookfield LV Viscometer) by selecting spindle number 64 and the viscosity of gels was noted [14].

### 5.3 Antimicrobial Study by Zone of Inhibition Method

Antimicrobial study was done as follows first the broth was prepared by adding yeast, peptone NaCl in water and then take 20 ml of volume into four separate boiling tubes for sub culturing the bacteria. In the next step nutrient broth was prepared by adding 28 grams of nutrient agar in 1000 ml of water. Then broth was sterilized and 20 ml of sterilized nutrient broth was taken in each petri dish i.e., four petridishes are taken for four bacterial (*staphylococcus aureus*, *E. coli*, *Proteous vulgaris*, *Bacillus*) inoculations. Sub cultured bacteria was spreaded on solidified medium. Then wells were prepared by using a sterile borer of diameter of 6 mm in petridish. Optimized celecoxib nanosponges gel, were added into each well separately [13]. Then the plates were incubated at 35–37 °C for 24 h. Then zone of inhibition of microbial growth around the well was measured in centimeters.

### 5.4 Ex Vivo Permeation Studies of F4 Loaded Celecoxib Nanosponges Gel

To find out the amount of drug permeated into the body, the *ex vivo* skin permeation studies was performed using goat skin, which was brought from the local slaughter house and it was kept in buffer pH 7.4 until it is mounted on Franz diffusion cell [10, 13]. Hair of the goat skin was removed by using a razor, wiped it with isopropyl alcohol. The skin membrane was placed in between two compartments (donor and receptor compartment). Receptor compartment is filled with phosphate buffer pH 7.4, in the donor compartment the skin is applied with the celecoxib loaded nanosponges gel [15, 16]. The cell was maintained at 37 °C and the stirring was kept at 100 RPM using a magnetic stirrer. Samples were taken at regular intervals of time 0, 0.25, 0.5, 1, 2, 3, 4, 6, 8, 12, 24 h and replaced with the freshly prepared buffer solution to maintain the sink condition and each sample was studied in triplicate and the average was taken into consideration for further calculation. The samples were analyzed by UV spectrophotometer (shimadzu, Japan) at 251 nm. Results were subjected for permeation parameters like permeation co-efficient, flux, lag time, diffusion co-efficient. The parameters were calculated by plotting a graph between percentage of CDR on the y-axis and time on x-axis. The data obtained in *ex vivo* study was subjected to mathematical models like, first order, zero order, Higuchi and korsmeyer-peppas model.

## 6 In Vivo Anti-Inflammatory Activity

Experimental animals Male Wistar albino rats were procured from Sri Venkateswara Enterprises, Bangalore were used in the study. The animals were housed under the standard environmental conditions like ambient temperature ( $25 \pm 1$  °C) humidity (55

$\pm 5\%$ ) and 12/12 h light dark cycles. All the experimental procedures and protocols which are used for the anti-inflammatory study was carried out in accordance with the guidelines of CPCSEA and which was approved by institution of animal ethical committee with Reg. No. 1677/PO/Re/S/2012/CPCSEA/IAEC/25 dt.3.05.2018. *In vivo* pharmacodynamic activity was performed to find out the efficacy of celecoxib loaded nanosponges gel in animals. For study male wistar albino rats were selected each weighing 150–200 gms and divided into five groups [17]. Depending on weight of rats dose was calculated. The inflammation was induced to hind paw by injecting carrageenan (1%w/v) using hind paw method. Rats were divided into five groups Group I Normal, Group II treated as control, Group III treated with marketed gel (Marketed Gel), [17] Group IV treated with pure celecoxib gel and Group V treated with celecoxib nanosponges gel. The paw volume was noted for every one hour using plethysmometer [18]. The percentage of edema inhibition was calculated by using the formula.

$$\% \text{ inhibition of edema} = \frac{(V_{\text{control}} - V_{\text{test}})}{V_{\text{control}}} \times 100 \quad (1)$$

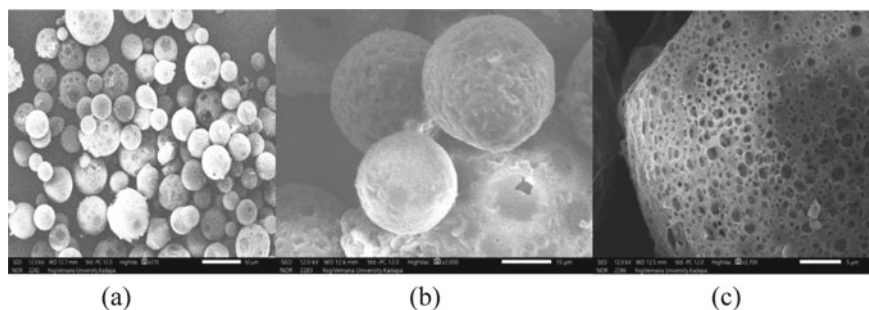
## 7 In Vivo Pharmacokinetic Studies

The pharmacokinetic study was performed to estimate the concentration of the drug in the blood for this the pharmacokinetic parameters were estimated by using HPLC (shimadzu, japan) method using water and acetonitrile (25:75) as mobile phase [19, 20]. Blood was withdrawn from the male wistar albino rats, sample was taken in ependroff tube at different time intervals, and sodium citrate was added to prevent coagulation. Then samples were centrifuged at 15,000 RPM for 30 min. The supernatant serum was collected. To it internal standard Flutamide (0.25  $\mu\text{g}$ ) was added [21]. 25  $\mu\text{l}$  of the sample was injected into Hplc column using Hamilton syringe and analyzed using RP-HPLC method at wave length 251 nm. (Group I control, Group II pure celecoxib, Group III celecoxib loaded nanosponges Gel).

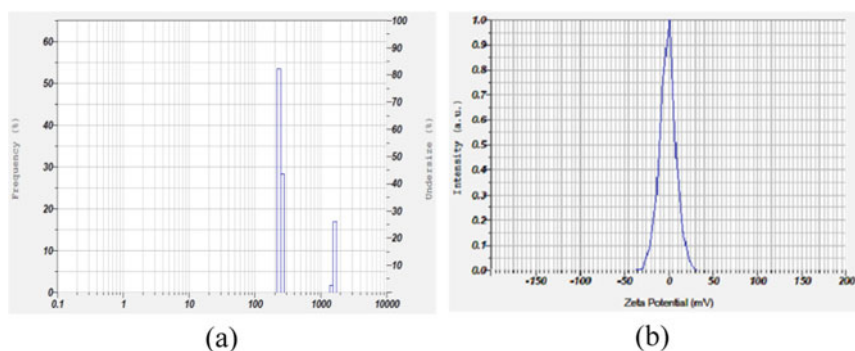
## 8 Results

### 8.1 SEM Analysis

The prepared nanosponges shape and surface morphology were analyzed by using scanning electron microscopy. They were found to be spherical in shape and its surface shows tiny mesh like structure. The SEM images of nanosponges are given in Fig. 1a, b and the surface morphology in Fig. 1c.



**Fig. 1** Scanning electron microscopy images (a–c)



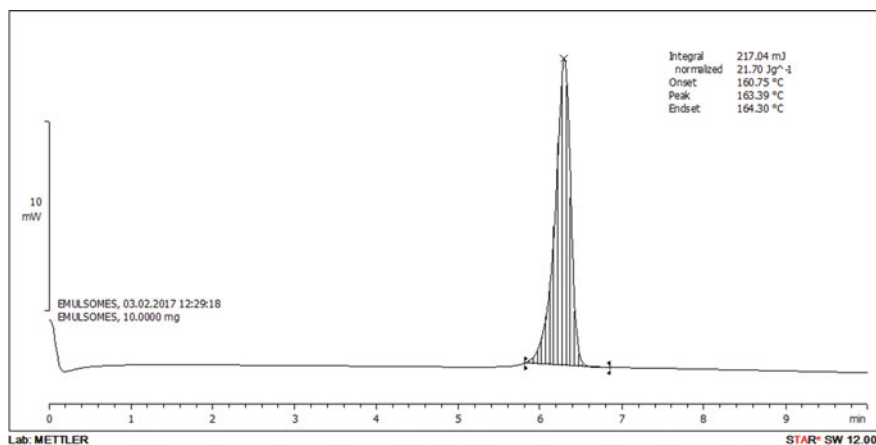
**Fig. 2** Zeta size and zeta potential of optimized F4 formulation of celecoxib loaded nanosponges

## 8.2 Particle Size and Zeta Potential

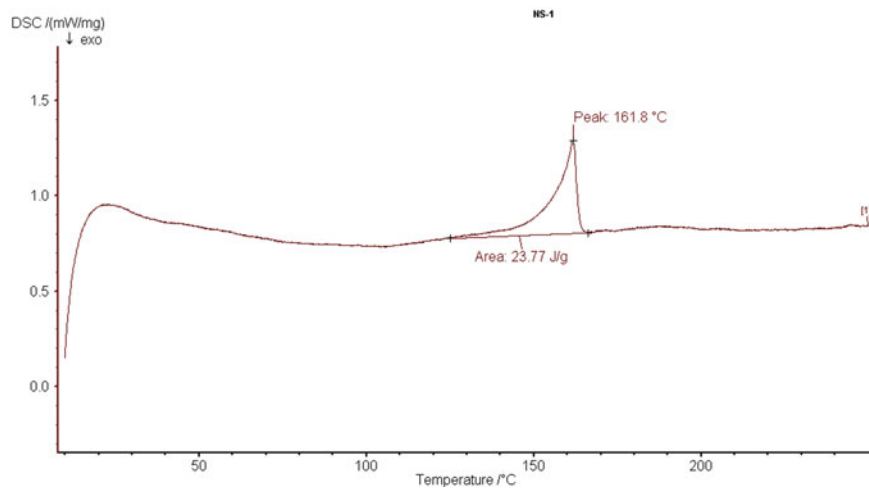
The zeta potential and particle size of the optimized nanosponges F4 formulation was analyzed using the zetasizer (Horiba), the samples were diluted in triplicates and the size of the nanosponges was found to be 240.9 nm for the optimized nanosponges (F4). The optimized F4 nanosponges zeta potential was found to be  $-1.8$  mV which shows that the formulation was stable. The zeta size and potential was shown in Fig. 2a, b.

## 8.3 DSC Analysis

The DSC thermogram of the pure celecoxib drug was obtained at  $163$  °C and in the celecoxib nanosponges formulation, it has shown slight decrease in the melting point at  $161$  °C. This implies that there is no chemical interaction between the drug and the polymer. The DSC thermograms were shown in the Fig. 3a, b.



(a pure celecoxib)

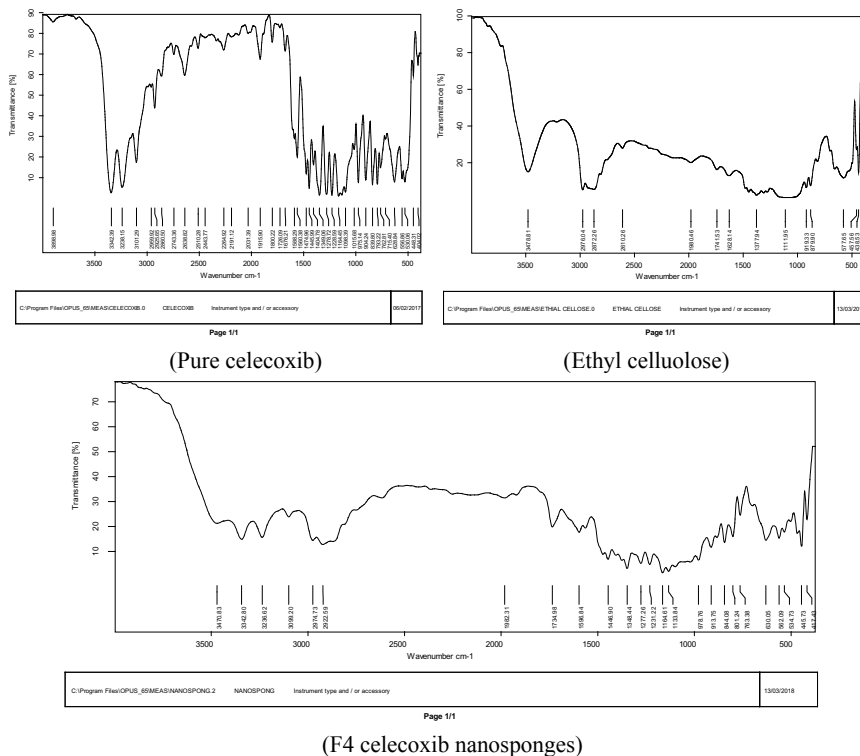


(b celecoxib nanosponges)

**Fig. 3** DSC thermograms of pure celecoxib and optimized F4 formulation of celecoxib loaded nanosponges

#### 8.4 Fourier—Transform Infra Red Spectroscopy

Fourier infrared spectroscopy shows that there is no chemical interaction between the drug and polymer in the prepared nanosponges, and all the characteristic peaks are present in the FT-IR spectrum of celecoxib loaded nanosponges which implies there is no chemical shift in the peaks when compared with the pure celecoxib drug. Celecoxib nanosponges showed its characteristic peaks at 3300–3500  $\text{cm}^{-1}$  which shows that they are related to N-H<sub>2</sub> stretching of sulfonamide groups. For S=O



**Fig. 4** Fourier transforms infra red spectroscopy of pure celecoxib and ethyl cellulose and optimized F4 celecoxib loaded nanosponges

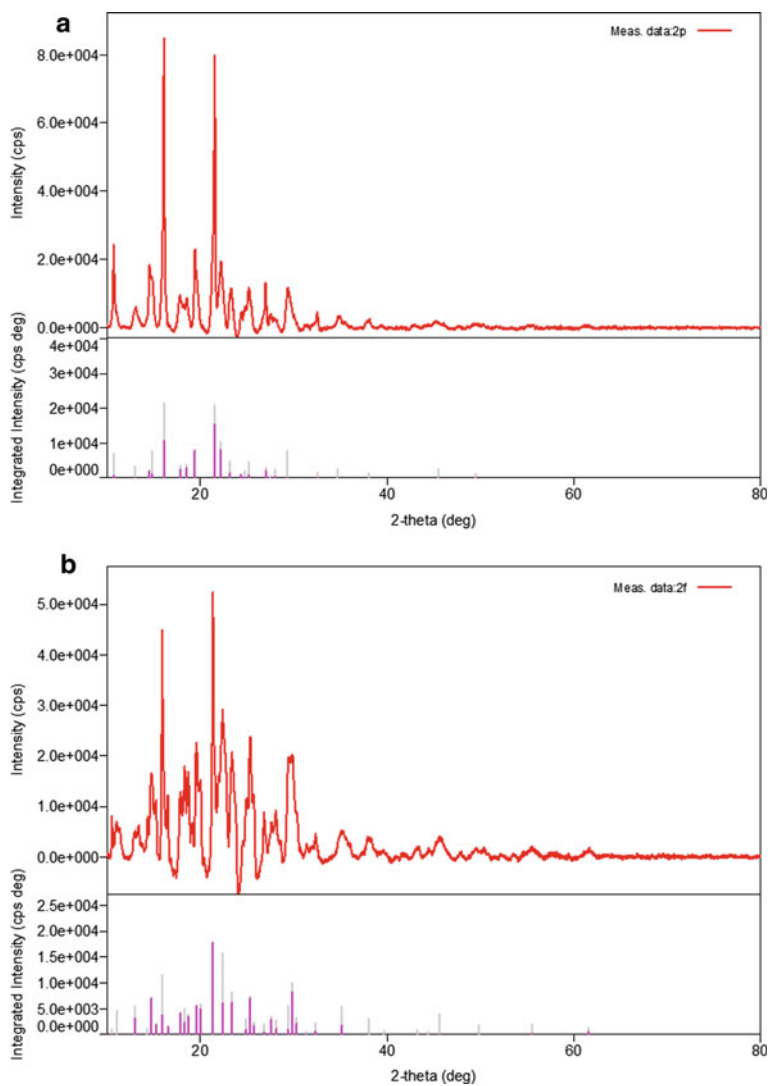
asymmetric and symmetric stretching group it is  $1150\text{--}1350\text{ cm}^{-1}$  and  $1550$  and  $1600\text{ cm}^{-1}$  for symmetric and asymmetric N-H stretching. The FT-IR spectrum of pure celecoxib and celecoxib loaded nanosponges were given the Fig. 4a–c.

## 8.5 X-ray Diffraction Studies

X-ray diffractograms of pure celecoxib, celecoxib nanosponges were recorded and the graph of them is as follows in the Fig. 5a, b.

## 8.6 Drug Entrapment Efficiency

Drug entrapment efficiency calculations were done for all the formulations and the highest amount of drug entrapped was found in F4 formulation and the amount of drug



**Fig. 5** X-ray diffractograms of pure celecoxib (a) and optimized F4 celecoxib loaded nanosponges (b)

entrapped was found to be 98%. The drug entrapment of the various formulations was tabulated in Table 3.

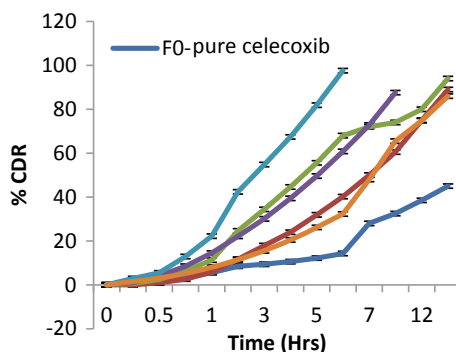
**Table 3** Drug entrapment efficiency of celecoxib loaded nanosponges

S.No	Formulation	Drug entrapment efficiency (%)
1	F1	96.5 ± 1.3
2	F2	96 ± 1.67
3	F3	97.2 ± 2.5
4	F4	95 ± 1.3
5	F5	98 ± 1.4
6	F6	23.5 ± 1.1
7	F7	8.5 ± 0.99
8	F8	13.4 ± 2.13
9	F9	8.5 ± 0.51

### 8.7 *In Vitro* Drug Release

All nanosponges formulations were performed to *in vitro* drug release studies, the cumulative percentages of drug release was obtained by taking a plot between the time on x-axis and cumulative percentage of drug release on y-axis. The studies were performed for all the formulations pure drug F0, F1, F2, F3, F4, F5. The amount of drug release was found to be 97.6% for F4 formulation in 6 h and for F3 98.1%, F2 89.2%, F1 80.6%, F5 86.02% and pure drug (F0) 45% at the end of 24 h they followed the zero order release profile. Cumulative percentage drug release of celecoxib nanosponges is given in the Fig. 6.

**Fig. 6** *In-vitro* drug release profile of celecoxib loaded nanosponges in pH 7.4 phosphate buffer from F0 (pure celecoxib)-F5 Formulations. The values are given as mean ± SD, n = 3





**Table 4** Evaluation of physicochemical properties of optimized X1 celecoxib loaded nanosponges gel

S. No.	Carbopol : HPMC	DRUG	Viscosity (cp)	Appearance	Ph
1	X1	HPMC: Carbopol (1:1)	8640	Homogenous	7.0 ± 0.2
2	X2	HPMC: Carbopol (2:1)	8820	Homogenous	7.0 ± 0.1
3	X3	HPMC:Carbopol (3:1)	9600	Homogenous	7.0 ± 0.2
4	X4	HPMC: Carbopol (4:1)	27,960	Homogenous	6.8 ± 0.3
5	X5	HPMC: Caroapol (5:1)	30,800	Homogenous	7.0 ± 1.2
6	X6	Celecoxib pure (1:1) Gel	8745	Homogenous	7.0 ± 0.53

The values are given as mean ± SD, n = 3

## 9 Physicochemical Evaluation of Celecoxib Loaded Nanosponges of Gels

### 9.1 PH

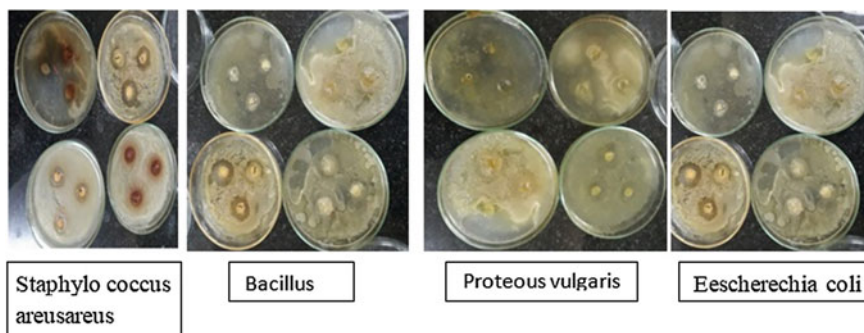
The pH for the formulations of gel X1–X6 showed  $7.0 \pm 0.2$  this indicates that the prepared gels were neutral and there is no skin irritation when applied to the skin. The pH of prepared nanosponges is shown in the Table 4.

### 9.2 Viscosity

The viscosity of the formulated gels was measured by using a Brookfield viscometer and the viscosity of the prepared gels X1–X6 was found to be in the range 8000–30,000 cps as the HPMC polymer concentration increases the viscosity of the gel has increased gradually. The viscosity of the optimized gel (X1) was found to be 8640cps. The viscosity of prepared nanosponges gel using different concentrations of carbapol943 and HPMC is shown in Table 4.

**Table 5** Zone of inhibition by bacterial strains in  $\text{cm}^2$  with mean  $\pm$  SD,  $n = 3$

Bacteria	Zone of inhibition ( $\text{cm}^2$ )
Staphylococcus aureus	2.34
Escherichia coli	2.80
Proteous vulgaris	3.2
Bacillus	3.0



**Fig. 7** Antimicrobial study by zone of Inhibition method for optimized nanosponges gel X1 loaded with F4 celecoxib nanosponges

### 9.3 Antimicrobial Study by Zone of Inhibition Method

In this study the bacterial strains *staphylococcus aureus*, *Escherichia coli*, *Proteous vulgaris*, *bacillus* were used [22]. The study was done for Celecoxib loaded nanosponges gel. The gels showed they were microbiologically inert towards the bacterial strains. The zone of inhibition was given in Table 5 and Fig. 7.

### 9.4 Ex-Vivo Skin Permeation Studies

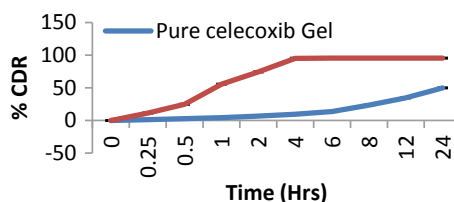
Optimized celecoxib loaded nanosponges gel X1 and pure celecoxib gel were subjected to *ex vivo* permeation studies using goat skin, which is placed in between the cells (Franz diffusion cell). The drug in the gel permeated through the goat skin into the receptor compartment was analyzed by taking the samples at regular intervals of time. The cumulative percent of drug permeated through the skin for optimized X1 gel loaded with F4 formulation was found to be 95% for four hours and for the pure celecoxib gel it showed 50% for 24 h. The cumulative percent drug diffused was given in Fig. 6. The parameters were calculated from the graph by plotting the cumulative percent drug release on y-axis versus time on x-axis. The flux was calculated from the slope of the graph and was found to be  $23.24 \mu\text{g}/\text{cm}^2/\text{hr}$  for celecoxib nanosponges and  $21.60 \mu\text{g}/\text{cm}^2/\text{hr}$  for pure celecoxib and the flux values was more

**Table 6** *Ex-vivo* skin permeation parameters of celecoxib nanosponges gel X1 loaded with F4 nanosponges and simple celecoxib gel

Formulations	Permeation co-efficient	Flux ( $\mu\text{g}/\text{cm}^2/\text{hr}$ )	Lag time	Diffusion coefficient
Celecoxib nanosponges gel	$2.81 \pm 0.3$	$23.24 \pm 2.1$	$0.581 \pm 0.002$	$0.0968 \pm 0.001$
Simple celecoxib gel	$2.16 \pm 0.1$	$21.60 \pm 3.2$	$1.004 \pm 0.0031$	$0.1676 \pm 0.0034$

The values are given as mean  $\pm$  SD, n = 3

**Fig. 8** *Ex vivo* skin permeation studies of pure celecoxib gel, celecoxib loaded nanosponges gel X1. The values are given as mean  $\pm$  SD, n = 3



for celecoxib nanosponges than pure celecoxib gel. The lag time, permeation coefficient, diffusion coefficient was calculated for celecoxib nanosponges gel and pure celecoxib gel and the values are as follows for celecoxib loaded nanosponges and pure celecoxib gel as shown in the Table 6. From the Fig. 8 it was found that the nanosponges loaded celecoxib gel follows zero order release mechanism, which governs the release. From equation  $R^2 = 0.99$  the drug release follows Higuchi model which indicates it follows the diffusion mechanism (Fickian  $n < 0.5$ ). The celecoxib loaded nanosponges gel showed the best release when compared to pure celecoxib gel.

## 10 Pharmacodynamic Studies

The pharmacodynamic activity of a pure celecoxib gel, celecoxib nanosponges gel, and marketed emulgel (voveran) was carried out using male Wister albino rats. The percentage inhibition of paw volume was calculated, it was found that percentage inhibition of edema were found to be, 76.2% for Group III rats, 57.6% for Group IV rats and 98.3% of Group V rats which is given in the Table 7. The results shows that there is a significant difference of ( $p < 0.05$ ) between pure celecoxib and celecoxib loaded nanosponges gel. Celecoxib loaded nanosponges and marketed gel shows effective inhibition when compared to pure celecoxib gel (celecoxib nanosponges gel > Marketed gel > pure celecoxib gel).

**Table 7** Grouping, treatment and percentage of inhibition of paw volume of rats

S. No.	Group	Treatment	% Inhibition (%)
1	Normal	Water	0
2	Group II (Control)	Only inducer	0
3	Group III	Marketed gel	76.27
4	Group IV	Pure celecoxib gel	57.6
5	Group V	Celecoxib nanosponges gel	98.3

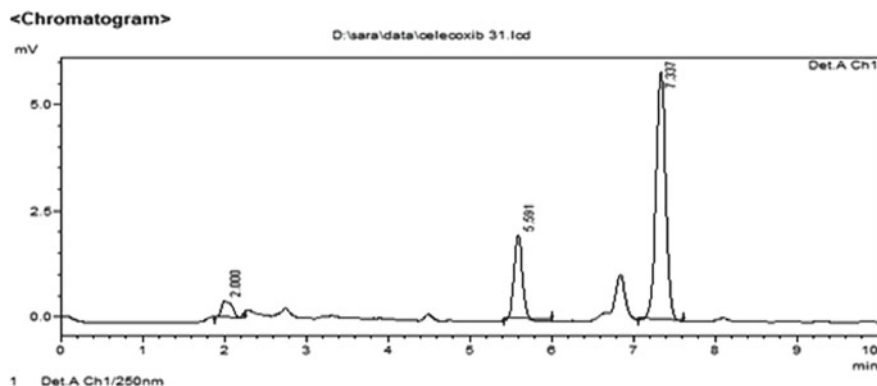
The values are given as mean  $\pm$  SD, n = 3

## 11 Pharmacokinetic Activity

The pharmacokinetic parameters for gel containing pure celecoxib and gels containing nanosponges of celecoxib were calculated using plasma concentrations. It was found that pure celecoxib gel has  $C_{max}$ ,  $T_{max}$ ,  $t_{1/2}$ , AUC as follows 23.13 mcg/ml, 8 h and 9.4 h, 90  $\mu$ g.hr/ml. Gel containing nanosponges of celecoxib has  $C_{max}$ ,  $T_{max}$ ,  $t_{1/2}$ , AUC was found to be 20.5  $\mu$ g/ml, 4 h, 8.5 h and AUC as 150  $\mu$ g/ml. This is given in Table 8 and Fig. 9. HPLC chromatogram of celecoxib in rat plasma after topical application of gel loaded with celecoxib nanosponges is shown in Figs. 9 and 10. The celecoxib loaded nanosponges gel has shown higher AUC values 150  $\mu$ g/ml and fast absorption with a  $C_{max}$  23.13  $\mu$ g/ml as it contains small, spherical size of nanosponges (240.9 nm) which allows the drug more permeable through the skin when compared to the pure celecoxib gel. The parameters (AUC),  $C_{max}$ ,  $T_{max}$ ,  $t_{1/2}$  were statistically significant ( $p < 0.05$ ) when compared to the pure celecoxib gel and celecoxib loaded nanosponges gel.

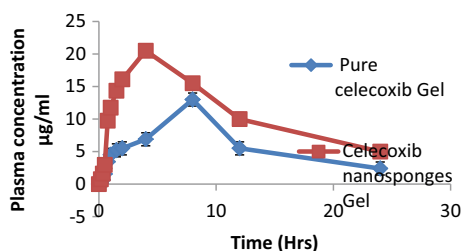
**Table 8** Pharmacokinetic parameters of pure celecoxib gel and celecoxib loaded nanosponges gel

Pharmacokinetic	Celecoxib pure gel	Celecoxib parameter loaded nanosponges gel
$C_{max}$	23.13 mcg/ml	20.5 mcg/ml
$T_{max}$	8 h	4 h
$t_{1/2}$	9.4 h	8.5 h
AUC	90 $\mu$ g.hr/ml	15090 $\mu$ g.hr/ml



**Fig. 9** HPLC chromatogram of celecoxib in rat plasma after topical application of gel loaded with celecoxib nanosponges gel X1

**Fig. 10** Plasma concentration- time profile of pure celecoxib gel and optimized celecoxib nanosponges gel X1



## 12 Discussion

The celecoxib loaded nanosponges were prepared by method emulsion solvent diffusion method using celecoxib and ethyl cellulose. The formed optimized nanosponges were found to be spherical in shape, with tiny mesh-like structure with a size of 240.9 nm and zeta potential  $-1.8$  mV. The sizes of the celecoxib loaded nanosponges from F1-F5 are as follows 50, 119.6, 124.1, 240.9, 496.5 nm. By particle size analysis we can conclude that the concentration of polymer shows its effect on the particle size. Optimized nanosponges F4 was found to have zeta potential  $-1.8$  mV which indicates that the particles are stable. Entrapment efficiency range was found to be in the range of 90 to 98.4%. FTIR studies show that there is no chemical shift between the peaks of celecoxib nanosponges when compared with the pure celecoxib drug. This shows that the polymer ethyl cellulose and drug celecoxib were inert without showing any interactions. DSC thermograms of pure celecoxib show the peak at  $163$  °C and that of the formulation at  $161$  °C this indicates that there is no incompatibility between the drug and polymer. To confirm the crystalline state celecoxib, celecoxib nanosponges powder XRD was performed for both. The powder XRD pattern of pure celecoxib showed sharp and significant, intense peaks are

observed at  $2\theta$  values of 10, 16, 21, 29 which are the characteristic nature of pure celecoxib. The diffractograms which are shown in the Fig. 5a, b suggested that there is no significant difference in peaks between the celecoxib nanosponges and celecoxib pure. The only slight difference in the intensity peaks in diffractograms between pure celecoxib and celecoxib nanosponges was because due to reduction in size of crystals, this reduction in size indicates the improvement in dissolution rate and its bioavailability. The *in vitro* drug release studies reveal that as the size of the nanoparticles shows 240.9 nm, it is easy to pass through the membrane and more amount of drug was also solubilized which showed *in vitro* drug release of 95% at 6 h and follows the zero order drug release. The optimized F4 formulation is converted into the gel using the polymers carbopol 934 and HPMC as the concentration of the carbopol increases the viscosity increases and after certain concentration the viscosity becomes more and aggregates are formed, and the stirring rate also plays an important role and it is observed that above 1000 rpm the breakage of gels is observed. Hence, for the preparation of gels, the parameters were optimized at 1000 RPM for 30 min the pH of the gel was adjusted to 7.0 by adding tri ethanol amine. The polymer carbopol 934 was selected because of its good bio adhesive property as it remains on the skin for longer periods of time and drug leakage was also minimum due to carbopol934 gel high viscosity. The percentage drug diffusion of celecoxib nanosponges gel was found to be 95% for 4 h as the solubility increases, more amount of drug will be permitted easily through the skin and thus the amount of drug release was more (4 h 95.5%) compared to pure celecoxib gel and a good linearity was observed with zero order release and the slope of the Higuchi model shows that it follows the diffusion pattern of drug release and the value indicates that it follows the fickian diffusion mechanism. As celecoxib belongs to BCS class II drug, it has less solubility and more highly permeation capacity. The drug permeation through stratum corneum of the skin can be done by two mechanisms one is by augmentation of skin permeability and secondly by activation of concentration independent transport driving system, and tri ethanol amine acts as a penetration enhancers, which increases the permeation of drug through the skin. Celecoxib loaded nanosponges posses better stability which was evident from pH, drug content and drug release from the stability studies. From the *in vivo* pharmacodynamic study, the percentage inhibition of edema in Group I, II rats are 0% as it is not treated. The percentage inhibition for group III and Group V rats are more 76.2 and 98.3% compared to Group IV 57.6% because Group III rats are treated with marketed gel and Group V rats were treated with celecoxib nanosponges gel as the solubility increases, permeation also increases thus it shows effective inhibition of edema compared to Group IV and Group III rats which were treated with pure celecoxib gel and marketed gel. The data were subjected to ANOVA, it was found that F ratio was 3.16 ( $f_c < f_t$ ). So there is a significant difference between the pure celecoxib and celecoxib loaded nanosponges gel. From pharmacokinetic parameters the *in vivo* bioavailability of celecoxib nano gel was found that  $C_{max}$ ,  $T_{max}$ ,  $t_{1/2}$ , AUC 20.5mcg/ml, 4 h, 8.5 h and AUC as 150 mcg/ml. As the nanosponges have more permeation the amount of the drug enters into blood circulation was more compared to pure celecoxib gel.

## 13 Conclusion

From above all the results it is concluded that nanosponges prepared with polymer ethyl cellulose and drug celecoxib by emulsion solvent diffusion method were effectively incorporated into the topical gel. The celecoxib nanosponges gel showed therapeutically effective treatment for anti-inflammatory activity compared to pure celecoxib gel and marketed voveran Emulgel.

### Declaration

**Conflicts of interest:** The authors declare that there are no conflicts of interest. This present research has not received any specific grant from any agency in public, commercial, not for profit sectors.

## References

1. P. Srinivas, A. Reddy, Formulation and evaluation of isoniazid loaded nanosponges for topical delivery. *Pharm Nanotechnol.* **3**(1), 68–76 (2015)
2. P.B. Subhashand, S.K. Mohite, Formulation Design & Development Of Artesunate Nanosponge. *Eur. J. Pharm. Med. Res.* **3**(5) (2016)
3. M. Cao, L. Ren, G. Chen, Formulation optimization and Ex vivo and In vivo evaluation of celecoxib microemulsion-based gel for transdermal delivery. *AAPS Pharm. Sci. Tech.* **2007** **18**(6), 1960–1971 (2008)
4. R. Bettini, P.L. Catellani, P. Santi, G. Massimo, N.A. Peppas, P. Colombo, Translocation of drug particles in HPMC matrix gel layer: effect of drug solubility and influence on release rate. *J. Controlled Release Official J. Controlled Release Soc.* **70**(3), 383–391 (2001)
5. G. Agarwal, M. Nagapal, G. Kaur, Development and comparison of nanosponge and noisome based gel for topical delivery of tazarotene. *Pharm. Nanotechnol.* **4**, 213–228 (2016)
6. Y. Liu, C. Sun, Y. Hao, T. Jiang, L. Zheng, S. Wang, Mechanism of dissolution enhancement and bioavailability of poorly water soluble celecoxib by preparing stable amorphous nanoparticles. *J. Pharmacy Pharm. Sci.* **13**(4), 589–606 (2010)
7. S. Ugursalgin, Ozgunvantanser: SYNTHESIS and characterization of beta cyclodextrin nanosponge and its application for removal of p-nitro phenol from water. *Clean Soil Air water* **45**(10), 2015
8. J. Patel, J. Trivedi, D.S. Chudhary, Formulation and evaluation of diacerein emulgel for psoriatic arthrities. *Int. J. Pharm. Res. Bio-Sci.* **3**(14), 625–638 (2014)
9. S. Arvapally, M. Harini, G. Harshita, Formulation and invitro evaluation of glipizide nanosponges. *Am. J. Pharm. Tech. Res.* **7**(3) (2017)
10. B. Mohanthy, K. Dipak, S. Mujumdar, A. Mishra K. Panda, Development and characterization of itracanazole loaded solid lipid nanoparticles. *Pharma Dev. Tech* **20**, 458–464 (2015)
11. S. Prathima Srenivas, Formulation and evaluation of voriconazole loaded nanosponges for oral and topical delivery. *Int. J. Drug Dev. Res.* **5**(1), 55–69 (2013)
12. M. Tuncay, S. Calis, H.S. Kas, M 2010: In vitro and in vivo evaluation of diclofenac sodium loaded album in microspheres. *J. Microencapsul.* **17**(2), 145–155 (2000)
13. E. Blanco-García, F.J. Otero-Espinar, J. Blanco-Méndez, J.M. Leiro-Vidal, A. Luzardo-Álvarez, Development and characterization of anti-inflammatory activity of curcumin-loaded biodegradable microspheres with potential use in intestinal inflammatory disorders. *Int. J. Pharm.* **518**(1–2), 86–104 (2017)
14. P. Karade, Formulation and evaluation of celecoxib gel. *J. Drug Deliv. Therapeutics* **2**(3), 132–135 (2012)

15. K.A. Anasari, P.R. Valli, F. Trota, R. Cavalla, Cyclodextrin based nanosponges for delivery of resebortol: invitro characterization, stability cytotoxicity and permeation study. *AAPS pharma Sci Tech.* **12**(1), 279–286 (2011)
16. M. Cao, L. Ren, G. Chen, Formulation optimization and Ex Vivo and In Vivo evaluation of celecoxib microemulsion-based gel for transdermal delivery. *AAPS Pharm Sci Tech.* **18**(6), 1960–1971 (2008)
17. S.H. Auda, S.A. El-Rasoul, M.M Ahmed, S.K. Osman, M. El-Badry, In-vitro release and in-vivo performance of tolmetin from different topical gel formulations. *J. Pharm. Invest.* **45**(3), 311–317 (2015)
18. E.A. Koçkaya, G. Selmanoğlu, K. Kismet, M.T. Akay, Pathological and biochemical effects of therapeutic and supratherapeutic doses of celecoxib in wistar albino male rats. *Drug Chem. Toxicol.* **33**(4), 410–414 (2010)
19. M.B. Majnooni, B. Mohammadi, R. Jalili, G.H. Bahrami, Rapid and sensitive high performance liquid chromatographic determination of zonisamide in human serum application to a pharmacokinetic study. *Indian J. Pharm. Sci.* **74**(4), 360–364 (2012)
20. S.K. Jain, M.K. Chourasia, R. Masuriha, V. Soni, A. Jain, N.K. Jain, Y. Gupta, Solid lipid nanoparticles bearing flurbiprofen for transdermal delivery. *Drug Delivery* **12**(4), 207–215 (2005)
21. S. Halirfroosh, K.O. West, Assesment of celecoxib poly (lactic-coglycolic) acid nanopformulation on drug pharmacodynamics and pharmacokinetics in rats european review for medicinal and pharmacological sciences, **20**(22) 4818–4829 (2016)
22. P. Srinivas, K. Sreeja, Formulation and evaluation of voricanazole loaded nanosponges for oral and topical deliver. *Int. J.drug Dev. Res.* (2015)



# Smart Bed Companion



G. V. V. S. Naveen, M. Shivani, Jalla Hasmitha, and D. Ajitha

**Abstract** After going through a busy and a strenuous day of life, all a person need is a night of good sleep. One needs nothing more but a comfy bed to rest and embrace into. When one has a night of sound sleep, the day starts fresh and blooming. One's mind finds itself at peace even in chaos. One thing that would help is a good bed. Currently in the market we only find quality cotton cots or made of some material that is suitable to sleep on according to our preference. Now in this smart world, what if that bed becomes smart? The aim of this paper is to manifest an idea of a smart bed companion that not only gives us good sleep but analyses sleep patterns and also controls many other applications and devices accordingly, also its unique factor (novel) to necessitate a person to wake-up. So this smart bed doesn't make us lazy but helps us to roll out on time.

**Keywords** Smart beds · Cots · Beds · Comfort · Adjustable beds · Sleep monitoring · Help sleep · Alarm · Internet of things · IoT in beds · Breathe control · Snoring · Snore preventing beds

## 1 Introduction

In this busy and hectic world, there should be some part of the day that gives us rest and keeps our mind fresh and relieves from all the stress. So, if you have good sleep you will have your mind fresh and concentrate on the work in our daily

---

G. V. V. S. Naveen · M. Shivani · J. Hasmitha (✉) · D. Ajitha  
Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, India  
e-mail: [hasmijalla@gmail.com](mailto:hasmijalla@gmail.com)

G. V. V. S. Naveen  
e-mail: [gvvsnaveen@gmail.com](mailto:gvvsnaveen@gmail.com)

M. Shivani  
e-mail: [shivanimanda99@gmail.com](mailto:shivanimanda99@gmail.com)

D. Ajitha  
e-mail: [ajithad@sreenidhi.edu.in](mailto:ajithad@sreenidhi.edu.in)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_47](https://doi.org/10.1007/978-3-030-46939-9_47)

life. Here comes the major problem, if you don't have proper sleep your day will be ruined. You get infuriated for the slightest troubles, exasperation, and unhealthy mind conditions are the leads. For people with busy schedules, the sleep they get for fewer hours should be at least peaceful and sound. People face many problems while sleeping, one reason may be the bed. They may not feel comfortable while sleeping. Sometimes they'll have body pains due to the texture, material or uneven surface of the bed. Sometimes the body posture doesn't go along with the bed. So, here comes our smart bed which is used for various purposes implemented by using the latest cutting edge technologies like IoT [1], Sensor networking, Cloud computing [1] etc.

## 2 Smartbed Features

### 2.1 What Our Smart Bed Does (Functionality)

**Posture alignment** According to our postures the bed gets adjusted within it. Our moves make the mattress align up and down to make ourselves comfortable.

**Prevent Snoring [3]** Some people have a snoring problem, due to which they may not breathe properly. Our smart bed adjusts the movement of the head in such a way that they can respire properly and prevent bad breath.

**Sleep monitoring [3]** It keeps track of the time you fall asleep, the time you wake up, the deep/light sleep patterns according to your respiration and body movements during sleep.

**Shake to wake** Bed will be attached with *thrusters*, so when you set an alarm to your bed, it starts vibrating slowly with slightly increasing intensity that creates 'have to get up' kind of mode (wake-up call). Also can have a music system player.

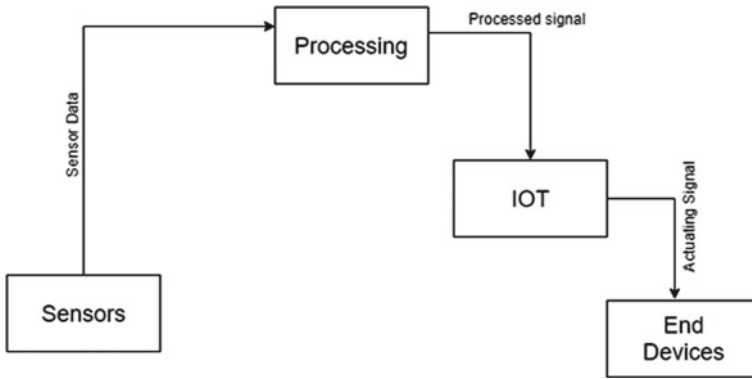
**Heartbeat monitoring** Pulse sensors embedded with the bed allow us to monitor the pulse rates and store it in the cloud. Whenever there's any unusual change in the pulse patterns, the user and the doctor are intimated immediately [4].

**Convertible** Bed to sofa or recliner.

**Temperature control** The bed becomes warm or cold according to the surrounding temperature or as set by the user [4].

### 2.2 Block Diagram

See Fig. 1.



**Fig. 1** Block diagram representing the flow of sensor data and the functioning of end devices accordingly

### 2.3 Description of the Block Diagram

**Sensors** Here we embed sleep monitoring sensors such as heartbeat sensor, BP sensor, sleep monitoring sensor, etc. in order to measure the various activities of the user and generate the corresponding information based on this which would be enabled for the user for reviewing [6].

**Sensor Data** This sensor data is the information gathered from the various sensors embedded into the cot and is sent for processing to generate the corresponding actuating signal [2].

**Processing** Here all the sensor data is fed as input to this section which consists of a Processor, which is the brain of this product, analyzes and maintains statistics of the records and sends out information to the IOT device [1].

**Processed Signal** The processed signal is the output signal generated from the processor unit which contains the information to be sent to different devices.

**IOT** Internet of Things enabled device is used in order to store the processed information in the cloud and provide remote access to the user regarding the sleep activities and provide a user-friendly interface for the user to control various IOT enabled devices in his house [2].

**Actuating signal** Actuating the signal is the one which triggers the respective end device based on the user’s choice [4].

**End Devices** These are the devices which are connected to the cot and get activated by the cot, this includes any household appliance such as Coffee maker machine, Geyser etc. this is also an interface for the user to monitor his sleeping activities and assign various input parameters for the device [5].

### 3 Areas of Application

Smart beds find its uses widely in various applications such as in residences, hospitals, dormitories, hostels, resting places, saunas, hotels, nursing homes etc.,. It can be an aid to elderly patients, physically challenged patients to help them concentrate on their pressure points during sleep. Smart bed intervenes the sleeping activity of normal persons by producing triggering signals so that the bed doesn't make them lazy. It also serves as a help to elderly patients with debilitated mobility.

### 4 Future Scope

Technology never stops getting bigger and better. As the technology advances, Smart bed will also have its upgradations accordingly. As of today, we see the IoT wearables being the boom of the market. Even a lot of public awareness is created of the advantages of using them. So people are also becoming smart along with the smart devices. They expect something new and better. Smart bed companion can be evolved as that smart device that interests a lot of people and also creates value. As a bed or a couch being a basic necessity to everyone, it will always remain in the trend. Aesthetic designs and multiple uses will stand out as unique attributes.

### References

1. M.M. Elsokah, A.R. Zerek, Next generation of medical care bed with internet of things solutions. In: 2019 19th international conference on sciences and techniques of automatic control and computer engineering (STA) 2019 Mar 24, pp 84–89 (2019). IEEE
2. R. Mieronkoski, I. Azimi, A.M. Rahmani, R. Aantaa, V. Terävä, P. Liljeberg, S. Salanterä, The internet of things for basic nursing care—A scoping review. *Int. J. Nurs Stud.* **1**(69), 78–90 (2017)
3. N.P. Patel, M.A. Grandner, D. Xie, C.C. Branas, N. Gooneratne, “Sleep disparity” in the population: poor sleep quality is strongly associated with poverty and ethnicity. *BMC Public Health* **10**(1), 475 (2010)
4. J. Kim, Analysis of health consumers' behavior using self-tracker for activity, sleep, and diet. *Telemedicine e-Health* **20**(6), 552–558 (2014)
5. J. Kim, A qualitative analysis of user experiences with a self-tracker for activity, sleep, and diet. *Interact J. Med. Res.* **3**(1), e8 (2014)
6. D. Poyares, C. Hirotsu, S. Tufik, Fitness tracker to assess sleep: beyond the market. (2015)

# Onion Husk Powder as a Adsorbent for Removal of Methylene Blue and Malachite Green from Aqueous Solutions



R. Usha, Ch. Indhravathi, D. Hymavathi, and M. Vijayalakshmi

**Abstract** A large quantity of synthetic dyes used for textile, industrial operations and releases highly harmful colored liquid in the form of wastewater. The effluents which are released from textile industries are becomes harmful to human beings as well as environment. The effluent must be treated before discharge into environment. Conventional and non-conventional treatment processes are frequently used to treat effluents. In the present study, we report experimental work based on adsorption of synthetic dyes—methylene blue and malachite green using onion husk powder. Maximum dye uptake capacity of methylene blue is 75.34% occurred at a 20 ppm of MB dye concentration, early MB pH of 7.0, onion husk powder dosage of 1 g/50 mL, reaction temperature of 303 K and reaction time of 90 min. Similarly, maximum uptake capacity of malachite green is 62.97% occurred at a 20 mg/L of dye concentration, an initial pH of 4.0, temperature 303 K, 0.5 g/50 mL of sorbent dosage and contact time of 90 min. Without any treatment of onion husk powder used as a sorbent for adsorption of synthetic dyes- methylene blue and malachite green up to 75.34 and 62.97%.

**Keywords** Dye removal · Malachite green · Methylene blue · Onion husk powder

## 1 Introduction

A numerous industries—textile, dye, plastic, dyestuffs, cosmetics, leather, pharmaceuticals and food [1, 2] emits effluents contains methylene blue and malachite green dyes. Removal dye particles became an important due to environmental issue and also harmful effects to living organisms and water ecosystem. The conventional

---

R. Usha (✉) · Ch. Indhravathi · M. Vijayalakshmi  
Department of Biotechnology, Sri Padmavati Mahila Visvavidyalayam (Women's University),  
Tirupati, Andhra Pradesh, India  
e-mail: [ushatirupathi@gmail.com](mailto:ushatirupathi@gmail.com)

D. Hymavathi  
Department of Chemical Engineering, S.V. University College of Engineering,  
Tirupati, Andhra Pradesh, India

methods—coagulation, photodegradation, ozonation and adsorption [3, 4] are used for the removal of different types of dyes and heavy metals from synthetic solution. The non-conventional methods—Fenton's reagent treatment and membrane filtration are not efficient for all types of dyes. Adsorption process is commonly used technique for the adsorption of dyes and heavy metals due to simple in operation [5]. Adsorbent, activated carbon is very effective for the adsorption of dyes and metals from wastewater through adsorption process [6, 7], but its high cost has call forth the research for adsorbents with low cost and easily available.

In present study, onion husk powder is used as a sorbent for adsorption of 1 of Methylene blue and Malachite green from synthetic solution.

## 2 Materials and Methods

### Preparation of adsorbent Onion husk powder:

Onion husk are collected in University Campus hostel, Tirupati, water washed neatly then doubled distilled water to eliminate surface impurities and sundried. The dried leaves are grounded and analysed with sieve shaker. The standard sieves of different mesh sizes—80, 100, 120 and 200 (177, 149, 125 and 75  $\mu\text{m}$ ) are used and the fractions are collected into separate bottles for experimental use [8].

### Stock solution preparation of Dye (Methylene blue/Malachite green):

The dye stock solution is prepared by a quantity, 1 g equivalent weight of dye and makeup with distilled water in a 1 L of standard flask (1000 mg/L). The experimental solutions of desired concentrations can prepared, by diluting the stock solution with distilled water. Solution pH can be adjusted with 0.1 acid/base solutions [7].

### Adsorption study:

A labelled number of Erlenmeyer flasks, each flask containing 50 mL solution with 50 mg/L of dye is taken and dye solution pH is adjusted with 0.1 N acid/base solutions. Acknowledged quantity of onion husk powder is added and the flasks are agitated at constant speed on platform shaker and at ambient temperature. Flasks are taken at suitable time intervals, the solution is filtered and the filtrate analysed using U.V Spectrophotometry, the method is continuous with different amounts of onion husk powder and other variables to make the study complete. Percentage adsorption of dye is [7, 8].

$$\% \text{Removal} = \frac{(c_0 - c_e)}{c_0} \times 100$$

where  $C_0$  early dye solution concentration, mg/L,  $C_e$  after dye solution concentration, mg/L.

### 3 Results and Discussion

The effects of various process variables are early dye concentration, dye solution pH, onion husk powder dosage and solution temperature on adsorption are studied and estimated. The experimental data are obtained under batch operation.

#### Effect of time on Adsorption

A 75  $\mu\text{m}$  average particle diameter of is used in the present study, pilot results indicated that a 90 min of reaction time is necessary for obtaining equilibrium, as shown in Fig. 1. Despite of size of particle and contact time, other four process variables - the dye molecule concentration (10–30) mg/L, early solution pH (2–8), onion husk powder quantity (0.5–1) g/L, and at room temperature, pressure the adsorption of dye in a highly interactive manner [9, 10].

A sample containing 20 mg/L of dye solution is reacted with 1 g/50 mL of onion husk powder for methylene blue and 0.5 g/50 mL of adsorbent for malachite green, at a pH of 7.0 for methylene blue and pH 4 for malachite green and at room temperature (303 K), yielded the maximum removal MB and MG are 75.34% and 62.5%. The effect of mixing time on adsorption of MB and MG is in Fig. 1.

From Figs. 1 depicts that. % removal of both MB and MG dyes increases with rise in time until equilibrium reached.

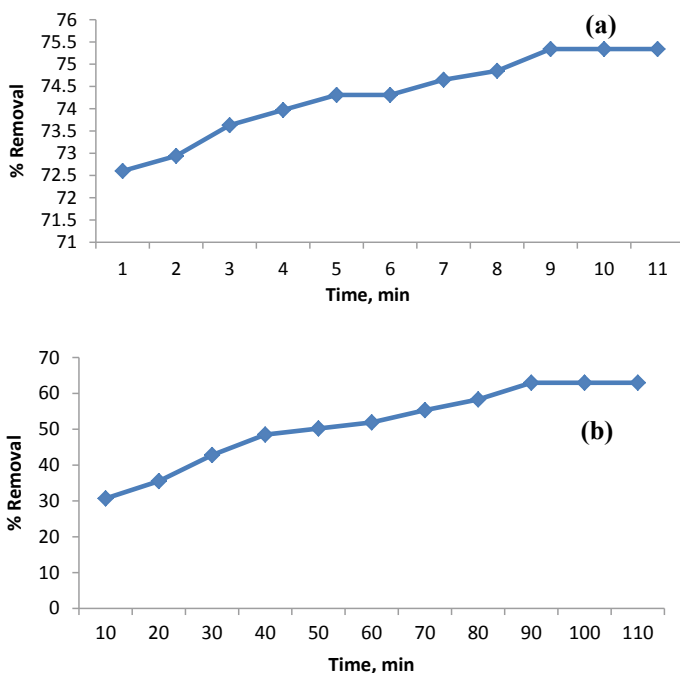
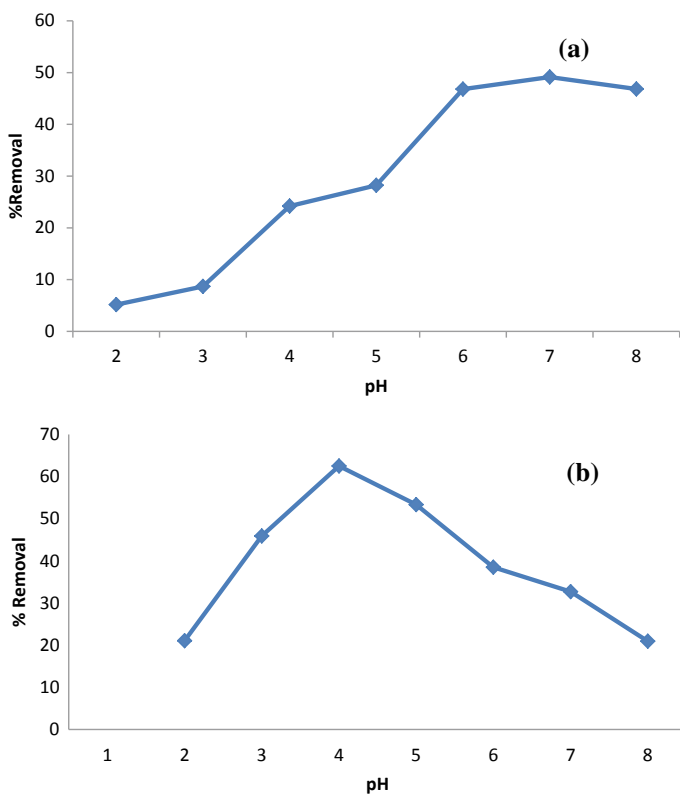


Fig. 1 Effect of mixing time on percentage removal of dye a methylene blue b malachite green

### Study of solution pH:

The role of solution pH is important in the removal of dye, this is as Fig. 2. It indicates the effect of solution pH on adsorption of MB onto the onion husk at early dye solution concentration of 20 mg/L, the quantity of onion husk powder 0.5 g/50 mL at 120 rpm agitation speed and at room temperature. Dye pH solution is changed from 2 to 8 by the addition of 0.1 N  $\text{H}_2\text{SO}_4$  or 0.1 N NaOH, the percentage adsorption of MB is raised from 5.17 to 49.15. In the present case, at low pH, the % adsorption of MB ions gets decreased and at higher pH, % removal of MB also increases up to pH of 7 and then decreases the % removal even rise in pH of the solution. As in the case of MB, percentage adsorption is raised from 21.05 to 62.15 with rises of pH from 2 to 4 and % adsorption decreases even increases of pH. In the present case, at acidic level, the adsorption of MB increased and pH of the solution changes from acid to base, % removal of MB is decreases from 62.5 to 30.95. The similar behaviour appears in the previous reports [10, 11].



**Fig. 2** Effect of pH on % removal by onion husk **a** methylene blue **b** malachite green



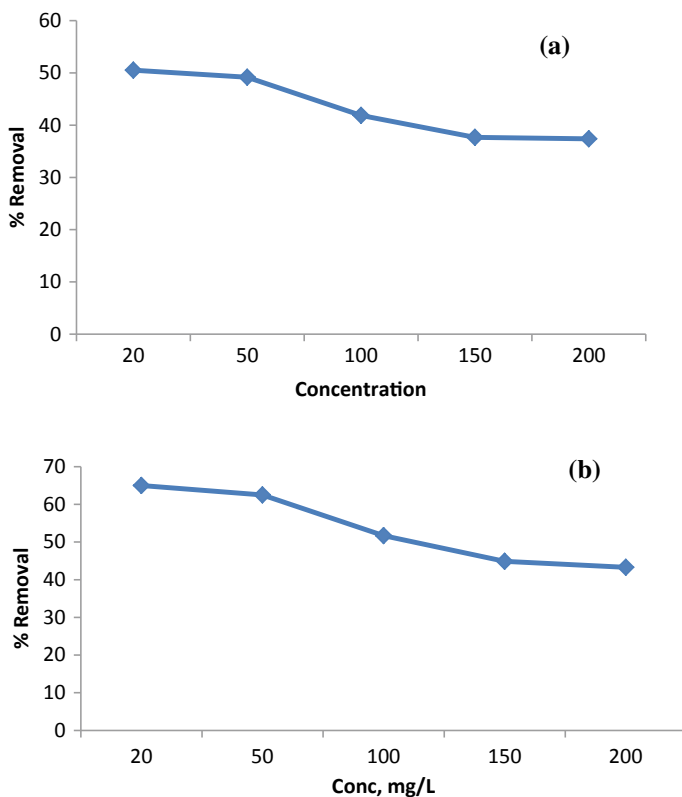
**Effect of dye concentration on Adsorption study:**

The variable solution concentration on the removal of dye plays an important role in adsorption process. The study of solution concentration of MB and MG onto onion husk powder is studied at different concentrations in Figs. 3a, b.

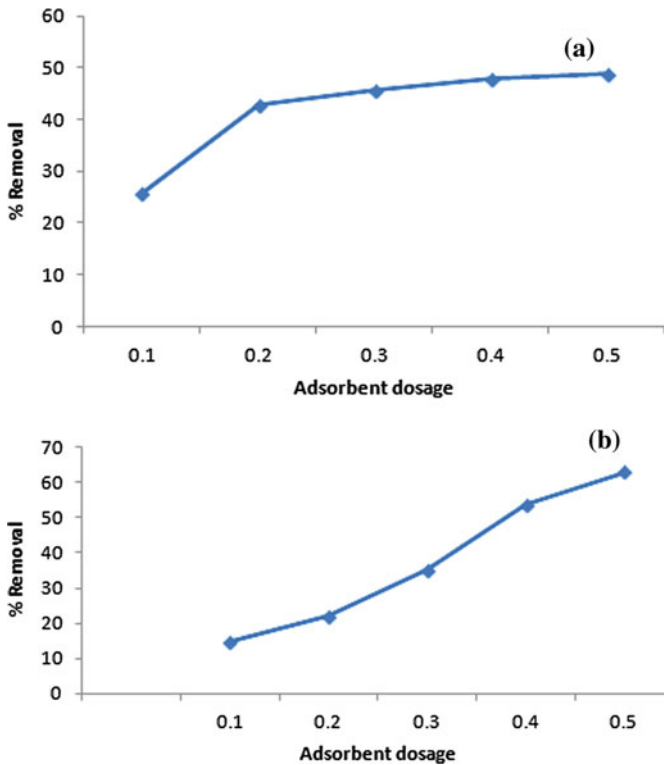
It is observed that the adsorption of MB decreases from 50.5 to 37.38 and for malachite green decreases from 60.5 to 43.3 with the rise in concentration from 20 to 200 ppm under the conditions are 90 min of equilibrium time, 20 ppm of early solution concentration, solution pH of 7.0 for MB and 4 for MG in Fig. 3b. Adsorbent dosage of 0.5 g/50 mL for both dyes, at room temperature [12].

**Effect of Dosage:**

Mixing is also crucial parameter in the sorption process, controls the distribution of the solute in the solution and formation of the boundary layer. The effect of adsorption of MB and MG using onion husk is considered at different dosages in Fig. 4. It indicates that the adsorption increases for both dye- MB and MG with increase in dosages of onion husk powder. Under these conditions are equilibrium



**Fig. 3** Study on early concentration on adsorption of dye **a** methylene blue **b** malachite green



**Fig. 4** Effect of adsorbent dosage on % removal **a** methylene blue **b** malachite green

time of 90 min, early concentration of 20 ppm, 0.5 g/50 mL onion husk powder, solution pH 7.0 for MB and 4.0 for MG at a room temperature.

From figures it reveals that the maximum uptake capacities of MB and MG are 75.34 and 62.97%.

The result is compared with other reports is good [13–15].

## 4 Conclusion

The onion husk powder is a potential adsorbent for removal of Methylene blue and Malachite green without any treatment of adsorbent. The investigation showed that the maximum uptake capacity of methylene blue and malachite green 75.34 and 65.5% obtained at a temperature of 303° K, a contact time of 90 min and 0.5 g/50 mL of adsorbent dosage, at a pH of 7.0 for MB and 4.0 for MG. Adsorbent, onion husk powder is cheap, easily available and environment-friendly.

## References

1. M.M. Hamed, I.M. Ahmed, S.S Metwally, Adsorptive removal of methylene blue as organic pollutant by marble dust as eco-friendly sorbent. *J. Ind. Eng. Chemi* **20**, 2370–2377 (2014)
2. M.T. Yagub, T.K. Sen, S. Afroze, H.M. Ang, Dye and its removal from aqueous solution by adsorption: A review. *Adv. Colloid Interface Sci.* **209**, 172–184 (2014)
3. Uma, B. Sushmitha, C.S. Yogesh, Equilibrium and kinetic studies for removal of malachite green from aqueous solution by a low cost activated carbon. *J. Indus. and Eng. Chemi.* **19**(4), 1099–1105 (2013)
4. D. Hymavathi, G. Prabhakar, Optimization, equilibrium and kinetic studies of adsorptive removal of cobalt(II) from aqueous solutions using *Cocos nucifera* L. *Chem. Eng. Comm.* **204**(9), 1094–1104 (2017)
5. Y. Fu, T. Viraraghavan, Fungal decolorization of dye wastewaters: a review. *Bioresourc. Technol.* **79**(3), 251–262 (2001)
6. A. Esra, A. Hüseyin, T. Mustafa, S. Ahmet, Effective removal of methylene blue from aqueous solutions using magnetic loaded activated carbon as novel adsorbent. *Chem. Eng. Res. and Design.* **122**, 151–163 (2017)
7. D. Hymavathi, G. Prabhakar, Studies on the removal of Cobalt(II) from aqueous solutions by adsorption with *Ficus benghalensis* leaf powder through response surface methodology. *Chem. Eng. Comm.* **204**(12), 1401–1411 (2017)
8. D. Hymavathi, G. Prabhakar, Biosorption of Pb(II) ions onto *cocos nucifera* leaf powder: application of response surface methodology. *Environ. Progress Sust. Energy.* <https://doi.org/10.1002/ep.12945>
9. N. Gupta, K.K. Atul, M.C. Chattopadhyaya, Application of potato (*Solanum tuberosum*) plant wastes for the removal of methylene blue and malachite green dye from aqueous solution. *Arabian J. Chemi.* **9**(1), S707–S716 (2016)
10. B.H. Hameed, A.A. Ahmad, Batch adsorption of methylene blue from aqueous solution by garlic peel, an agricultural waste biomass. *J. Hazard. Mater.* **164**, 870–875 (2009)
11. G.L. Dotto, M.L.G. Vieira, V.M. Esquerdo, L.A.A. Pinto, Equilibrium and thermodynamics of azo dyes biosorption onto *Spirulina platensis*. *Brazilian J. Chem. Eng.* **30**(1), 13–21 (2013)
12. G. Renmin, J. Youbin, C. Fayang, C. Jian, L. Zhili, Enhanced malachite green removal from aqueous solution by citric acid modified rice straw. *J. Hazard. Mater.* **137**, 865–870 (2006)
13. Z.M. Hussin, N. Talib, M. Hussin, A.K. Megat, M. Hanafiah, K. Wan, A.W.M. Khalir, Methylene blue adsorption onto NaOH modified durian leaf powder: isotherm and kinetic studies. *Am. J. of Environ. Eng.* **5**(3A), 38–43 (2015)
14. Z. Wexuan, Y. Han, J.L. Hai, J. Ziwen, D. Lei, K. Xiwei, Y. Hu, L. Aimin, C. Rongshi, Removal of dyes from aqueous solutions by straw based adsorbents: batch and column studies. *Chem. Eng. J.* **168**, 1120–1127 (2011)
15. K. Vasanth Kumar, Optimum sorption isotherm by linear and non-linear methods for malachite green onto lemon peel. *Dyes Pigm.* **74**(3), 595–597 (2007)

# Bioalgalization—A Novel Approach for Soil Amendment to Improve Fertility



Layam Anitha, Gannavarapu Sai Bramari, and Pilla Kalpana

**Abstract** Current problems of world including food security, water scarcity, soil erosion, climate changes, population demand and environmental safety can be challenged by agriculture science by introduction of biotech crops, new farming practices and new crop protection methods. The efficiency of crops is improved by a novel technique like bioalgalization for soil amendments. In this aspect, *Spirulina* is applied to soils along with biofertilizers, organic manure and vermicompost anticipating enhanced soil mineral status to help the growth and yield of crops. The present experiment was carried out with field studies on Amaranthus, Green gram and Tomato using different combinations and concentrations of *Spirulina* with biofertilizer, vermicompost and organic manure and different treatments to estimate the NPK status in plants and in soils prior and after the studies. There was 10–20 fold increase of protein content in yield of tomato when compared with reference value of 0.9/100 g with different concentrations of *Spirulina*. The soil nitrogen levels were found to be increased in experimental set up of green gram seeds soaked in *Spirulina*, 5 g concentration resulted in N content as  $(0.84 \pm 0.04\%)$  compared to control  $(0.03 \pm 0.02\%)$ . In experimental method of biofertilizer and *Spirulina* combination Phosphorus content of soil after harvest of *Amaranthus* plants was  $44.5 \pm 0.70$  mg/100 g and the control value was  $37 \pm 0.70$  mg/100 g. In post-harvest soil of tomato plants the potassium (K) levels were increased to  $184.5 \pm 2.1$  mg/100 g from the control value of  $44 \pm 0.70$  mg/100 g in 3 h of soaking experimental group. Bioalgalization is a promising technology to prevent soil erosion and pollution caused by use of heavy chemical fertilizers and also helps to improve soil fertility.

**Keywords** *Spirulina* · Biofertilizer · Vermicompost · Amaranthus · NPK status

---

L. Anitha (✉)  
Princess Nora University, Riyadh, Saudi Arabia  
e-mail: [layamanitha@gmail.com](mailto:layamanitha@gmail.com)

G. S. Bramari · P. Kalpana  
CHEGG India Pvt. Ltd, New Delhi, India

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_49](https://doi.org/10.1007/978-3-030-46939-9_49)

## 1 Introduction

Key issues such as water scarcity, soil erosion, climate changes, population demand and environmental safety are to be addressed by improving the agriculture technology, thereby improving crop yields for ensuring food security according to demand. The soil should be supplied with sufficient nutrients and is amended with effective fertilization techniques. Balanced fertilization ensures adequate availability of nutrients in soil, so as to meet the requirement of crops at critical stages of growth. The efficiency of crops is improved by reduction of usage of chemical fertilizers, practicing conservative agriculture farming which focus on reducing the adverse impacts on the environment by promotion and utilization of biofertilizers and organic manures globally [1].

The use of biofertilizers in current agricultural practices enhances the crop growth and yield. The biofertilizers are applied along with vermicompost, organic manure and microbial inoculants [2]. The microbial biofertilizers used include seaweed liquid extracts, microbial inoculants, biostimulators and biofortification agents. All these categories of microbial biofertilizers are involved in the enhancement of plant nutrient uptake. This in turn results in crops with high vitamin and nutrient content, and high yields [3]. Agriculture should play a major role in improving nutrition, in particular by paying more attention to the nutritional quality of food. Biofortification is a scientific method for improving the nutritional value of foods. Biofortification holds great promise for improving nutrition and should continue to be explored [4].

Biofortification has four main advantages when applied in the context of the poor in developing countries. First, it targets the poor who eat large amounts of staple foods daily. Second, biofortification targets rural areas where it is estimated that 75% of the poor live mostly as subsistent or small hold farmers, or landless labourers. These populations rely largely on cheaper and more widely available staple foods such as rice or maize for sustenance [23]. Biofortification is cost effective and has distinct advantages that can complement other traditional approaches to improve nutrition. Microbial inoculants have been the primary focus of programs to enhance staple food crops with sufficient levels of Fe, Zn, and provitamin A, carotenoids [5].

The utilization of bacterial, algal and fungal biofertilizers were recommended. In this aspect use of *Spirulina* application to soils along with biofertilizers or vermicompost is studied for increasing the levels of protein, macro and micronutrients in soil as well as in the yield. *Spirulina platensis* can be used as a microbial inoculant or a biofortification agent to the plants to enhance the nutrients and helps to improve the nutrient status of human population and in turn reducing malnutrition. Thus, the present study aims to appraise the soil fertility using *Spirulina* by pot studies. Field studies are carried out with different concentrations and fertilizer combination methods of application of *Spirulina* and study its effect on NPK status of crops and as well as soil.

## **2 Materials and Methods**

The experiment was carried out with different methods such as soaking of seeds in *Spirulina* extract for different time periods, different concentrations, and different proportions of *Spirulina* extract mixed with biofertilizers, vermicompost, organic manure and foliar application of *Spirulina* extract in different concentration. Further, nutrient profile of *Spirulina*, soil prior to experimentation and nutrient profile of the yield has been studied.

### ***2.1 Estimation of Nitrogen and Protein***

The Nitrogen and protein content was analyzed by Kjeldahl method by kjeltech model, Pelican Kelplus—KES 12 INL [6].

### ***2.2 Estimation of Phosphorus***

The amount of phosphorous was estimated by Fiske and Subba Row Calorimetric method [7].

### ***2.3 Estimation of Potassium and Sodium***

The amount of Potassium and Sodium was estimated by Elico CL 22D Flame Photometer [8].

### ***2.4 Estimation of Nitrogen, Phosphorous and Potassium in Soil***

The amount of N, P and K present in soil was estimated by Soil Testing Centre, Government Limited [9].

### ***2.5 Protein and Nitrogen Estimation in Yield of Crops***

Estimation of Protein Nitrogen in *Amaranthus*, Green gram and Tomato samples was carried out by Kjeldahl method Pelican Kelplus—KES 12 INL [6].

### 3 Results and Discussion

#### 3.1 Estimation of Soil Nutrient Profile

The soil nutrient profile was studied and the soil was analyzed for minerals like: Nitrogen, Phosphorous and Potassium. Studies were carried out on soil properties like pH and Electric conductivity. All nutrients were high in the soil treated with *Spirulina* when compared with control except potassium (Table 1).

The yield of the crop mainly depends on the soil health which is a crucial factor. The soil pH, conductivity and nitrogen have been reported higher when compared with control. Physical and chemical properties of soil from Olasati village were studied in detail and carried out the pot experiment analysis for response of green gram to the application of Minjingu Mazao fertilizer (31% P<sub>2</sub>O<sub>5</sub>) in Tanzania, and the results indicated that N, P, K in soil were increased [10]. Zodape et al., [11] studied the foliar application of seaweed extract. *Kappaphycus alvarezii* extract was applied on green gram plants (*Phaseolus radiata* L) to enhance the yield and nutritional quality. The yield was analyzed for total carbohydrates, total Nitrogen, total protein and various micro nutrients like Fe, Zn and Cu etc. The foliar application resulted in high yield and improved nutrients of green gram.

The field experiment was carried out to study the effect of chemical fertilizers, Biofertilizers, Organic matter, vermicompost combination with *Spirulina* and soaking in *Spirulina* hydrolysate to enhance the level of minerals like P, K, Zinc and Iron, the components like Nitrogen and Protein in *Amaranthus*, Green gram and Tomato plants. Due to the leaching of Nitrogen, phosphorous and potassium from the soil the amounts of these elements have to be replenished in soil with fertilizers. Chemical fertilizers like Urea, super phosphates and potassium fertilizers are widely used.

The Table 2 shows the effect of various experimental methods resulting in increased status of NPK levels in *Amaranthus* plants. Time period soaking is a traditional approach that allows the transport of mineral ions between *Spirulina* extract and *Amaranthus* seeds. Different concentrations method also works on same principle but the equilibrium of ions is established at a certain concentration. *Spirulina* in

**Table 1** Effect of *Spirulina* on selected soil nutrient profile of pot study prior and after experimentation

Parameters	Pre-experimentation	Post experimentation	Control
pH	7.0	7.3	7.1
E.C ds m <sup>-1</sup>	0.23	0.37	0.12
Nitrogen (%)	0.50	0.55	0.36
Phosphorous mg/100 g	33	39	37
Potassium mg/100 g	124	81	62
Zinc mg/100 g	26	32	28

**Table 2** Effect of *Spirulina* supplementation on *Amaranthus* plants

Experimental methods	Protein (%)	Nitrogen (%)	Phosphorus (mg/g)	Potassium (mg/g)
Time period soaking	9.8 ± 0.00 (4 h)	1.5 ± 0.00 (4 h)	332 ± 0.70 (4 h)	121 ± 0.70 (3 h)
Soaking in different concentration	9.5 ± 0.00 (15 g)	1.5 ± 0.00 (15 g)	332 ± 0.70 (30 g)	116.5 ± 0.70 (30 g)
Biofertilizer	13.0 ± 0.00 (25:75)	2.0 ± 0.00 (25:75)	356. ± 0.70 (75:25)	111.5 ± 0.70 (75:25)
Vermicompost	11.5 ± 0.00 (75:25)	2.6 ± 6.36 (75:25)	302 ± 2.10 (50:50)	126 ± 1.41 (50:50)
Organic manure	13.3 ± 0.00 (75:25)	2.1 ± 0.00 (75:25)	211 ± 1.41 (25:75)	106 ± 1.41 (50:50)
Chemical fertilizer	11.9 ± 0.00 (50:50)	1.9 ± 0.00 (50:50)	276 ± 0.70 (50:50)	103 ± 1.41 (25:75)
Spray method	12.9 ± 0.00 (25/5L)	2.07 ± 0.00 (25/5L)	228 ± 0.70 (25/5L)	173.5 ± 4.94 (25/5L)

combination with biofertilizer, vermicompost, organic manure and chemical fertilizer work at different proportions and is evident from the table. Foliar application is a novel method that directly interacts with the plant leaf cells allowing the entry of mineral ions improving their concentrations [12]. The combination of cyanobacteria and flyash has known to improve the total soil nitrogen status and as well as other mineral status in rice crops and soil in which they were grown [13].

Seed priming is one of the promising technologies that influence the plant—soil interactions, mineral status and microbial community colonization in root zones of plants. Seed bioprimering with *Spirulina* (soaking) has turned out to be novel method that improves soil as well as crop mineral characteristics. Therefore, seed priming enables to improve the soil mineral levels and thereby also help to increase crop growth and yield [19].

The Table 3 describes the effect of various experimental methods in improving the yield of green gram plants by utilization of *Spirulina*. High amounts of nitrogen (%N) were observed in experimental set up *Spirulina* + biofertilizer combination in 25:75 proportions. Increased amounts of phosphorus were observed in soaking method using 15 g of *Spirulina*. Potassium levels were found to be increased in the experimental method of *Spirulina* + chemical fertilizer in 25:75 proportions. The treatment of soil and chick pea crops with a combination of cyanobacteria and bacterial biofilms have enhanced various physical, biochemical and morphological parameters of the crops and also the mineral status of soil due to increased nitrogen availability in soil [14].

Soil microbial interactions play a major role in soil amendments that in turn increase the overall availability of soil minerals to crops due to which the crop



**Table 3** Effect of *Spirulina* supplementation on green gram plants

Experimental methods	Protein (%)	Nitrogen (%)	Phosphorus (mg/g)	Potassium (mg/g)
Time period soaking	22.7 ± 0.00 (4 h)	3.6 ± 7.07 (4 h)	331 ± 1.41 (3 h)	122.5 ± 0.70 (1 h)
Soaking in different concentration	21.7 ± 0.00 (10 g)	3.4 ± 7.07 (10 g)	347 ± 0.70 (15 g)	84.5 ± 0.70 (20 g)
S + Biofertilizer	25.5 ± 0.00 (25:75)	4.0 ± 7.07 (25:75)	253 ± 0.20 (75:25)	57 ± 1.41 (25:75)
S + Vermicompost	24.1 ± 0.00 (25:75)	3.8 ± 7.07 (25:75)	243 ± 2.12 (75:25)	74.5 ± 6.70 (75:25)
S + Organic manure	23.8 ± 0.00 (75:25)	3.8 ± 0.00 (75:25)	231 ± 0.70 (75:25)	63 ± 4.24 (50:50)
S + Chemical fertilizer	24.1 ± 0.00 (50:50)	3.8 ± 0.00 (50:50)	308 ± 1.41 (25:75)	88.5 ± 0.70 (25:75)
Spray method	21.3 ± 0.00 (25/5L)	3.4 ± 7.07 (25/5L)	217 ± 1.41 (50/5L)	85.5 ± 2.12 (50/5L)

growth and yield are increased. In this experimental method, the synergistic interactions between *Spirulina* and biofertilizers were thought to have an important role in increasing soil nitrogen, phosphorus and potassium status [18].

Table 4 depicts the effect of *Spirulina* on tomato plants using different experimental methods. In tomato plants the % of Nitrogen was found to be higher in *Spirulina* + organic manure method in 50:50 proportions. Phosphorus levels were higher in *Spirulina* + biofertilizer and *Spirulina* + vermicompost methods in 75:25 proportions. High potassium levels were found in 5 h soaking method. Increased qualitative and quantitative traits have been observed in cherry tomato plants. The plant seeds were

**Table 4** Effect of *Spirulina* supplementation on tomato plants

Experimental methods	Protein (%)	Nitrogen (%)	Phosphorus (mg/g)	Potassium (mg/g)
Time period soaking	19.9 ± 0.00 (5 h)	3.1 ± 7.07 (5 h)	210 ± 0.71 (O.N)	127 ± 1.41 (5 h)
Soaking in different concentration	21.0 ± 0.00 (20 g)	3.3 ± 7.07 (20 g)	206 ± 7.07 (30 g)	84.5 ± 0.70 (20 g)
S + Biofertilizer	22.2 ± 0.00 (25:75)	3.4 ± 7.07 (25:75)	332 ± 0.71 (75:25)	58.0 ± 1.41 (25:75)
S + Vermicompost	19.2 ± 7.07 (75:25)	3.0 ± 7.07 (75:25)	332 ± 0.71 (75:25)	66.5 ± 0.70 (50:50)
S + Organic manure	23.1 ± 7.07 (50:50)	3.6 ± 7.07 (50:50)	294 ± 6.36 (75:25)	63.0 ± 4.24 (50:50)
S + Chemical fertilizer	17.1 ± 7.07 (75:25)	2.7 ± 7.07 (75:25)	243 ± 2.12 (50:50)	78.5 ± 0.70 (25:75)
Spray method	20.6 ± 7.07 (100/5L)	3.3 ± 0.00 (100/5L)	227 ± 1.41 (50/5L)	85.5 ± 2.12 (50/5L)

**Table 5** Effect of *Spirulina* supplementation on % nitrogen in post-harvest soil of experimental plants

Experimental methods	Nitrogen (mg/g)		
	A	G	T
Time period soaking	0.77 ± 0.00 (4 h)	0.83 ± 0.042 (4 h)	0.80 ± 0.00(5 h)
Soaking in different concentration	0.80 ± 0.02 (15 g)	0.84 ± 0.04 (5 g)	0.68 ± 0.12 (5 g)
S + Biofertilizer	0.74 ± 0.04(25:75)	0.57 ± 0.07(50:50)	0.03 ± 0.02(50:50)
S + Vermicompost	0.81 ± 0.02 (50:50)	0.60 ± 0.02 (25:75)	0.60 ± 0.02(50:50)
S + Organic manure	0.76 ± 0.01(in control)	0.04 ± 0.01(25:75)	0.03 ± 0.02(in control)
S + Chemical fertilizer	0.83 ± 0.02(75:25)	0.03 ± 0.02(25:75)	0.03 ± 0.02(50:50)
Spray method	0.82 ± 0.00(50 g/5L)	0.04 ± 0.01(25 g/5L)	0.03 ± 0.01(25 g/5L)

A Amaranthus; G Green gram, T Tomato

treated with combination of brown algae *Ascophyllum nodosum* and vermicompost extracts. The experiment resulted in overall increase of all parameters of crops and the experiment was carried out in hydroponic conditions [15].

Integrated nutrient management of soil and crops is essential to improve the crop yield of tomato. Conventional agricultural methods are followed by application of biofertilizer and organic manures increased the plant morphological characters, yield and physiological status was improved by increased nutritional status. Utilization of effective soil amendment methods had resulted in increase in the yield of tomato in the current experiment [20]. The Table 5 shows increase in % Nitrogen in post harvest soil of three experimental plants. High amounts of Nitrogen were found in green gram plants in 5 g soaking method followed by Amaranthus plants in *Spirulina* + chemical fertilizer method in 75:25 proportions. The high amount of nitrogen in tomato plants was  $0.80 \pm 0.00$  and was reported in 5 h soaking method.

Soil mineral status is known to improve by crop—soil interactions along with soil amendments. The utilization of biofertilizers, organic manure and vermicopost increased soil fertility by increasing soil nitrogen status. Soil nitrogen is improved by reducing the soil acidity and also preserving the root zone microorganisms attracting more microbes with root exudates. Application of biofertilizers along with *Spirulina* helped increasing the protein content of crops and nitrogen levels of soil since, *Spirulina* is protein rich cyanobacterial member that obviously contributes more nitrogen to post harvest soil [21]. The N, P and K in soil were found to increase significantly due to inoculation of cyanobacteria. High NPK levels were observed when the seed soaking + soil drench +75% of N was applied to plants [24].

The Table 6 shows the post-harvest phosphorus levels in soils after growing the experimental plants *Amaranthus*, Green gram and Tomato. High amounts of phosphorus were reported in green gram plants in soaking in 25 g concentration method.

**Table 6** Effect of *Spirulina* supplementation on % phosphorous in post-harvest soil of experimental plants

Experimental methods	Phosphorous (mg/g)		
	A	G	T
Time period soaking	34.5 ± 0.70 (3 h)	45 ± 1.41 (3 h)	41.5 ± 2.12 (5 h)
Soaking in different concentration	44.5 ± 0.70 (30 g)	45.5 ± 2.12(25 g)	42 ± 2.82 (10 g)
S + Biofertilizer	44.5 ± 0.70(25:75)	42 ± 0.70 (in control)	38 ± 0.70 (in control)
S + Vermicompost	40.0 ± 0.70(in control)	43 ± 1.41 (50:50)	46 ± 1.41 (75:25)
S + Organic manure	44.5 ± 0.70 (50:50)	37 ± 1.41(50:50)	36 ± 0.70 (75:25)
S + Chemical fertilizer	36.5 ± 0.70 (25:75)	32 ± 0.70 (50:50)	36 ± 0.70 (in control)

A Amaranthus; G Green gram; T Tomato

The high amounts of phosphorus levels were followed in soils after harvesting Amaranthus plants. The high amounts were observed in experimental methods of soaking in 30 g concentration, *Spirulina* + biofertilizers in 25:75 proportions and *Spirulina* + organic manure in 50:50 proportions. Foliar application of *Spirulina* was proved to be equally effective as NPK fertilizers in enhancing soil NPK status. This can be evident by study of yield of eggplants using a foliar *Spirulina* fertilizer called Spirufert. Soil mineral status is also improved with supplementation of biofertilizers and organic manures that stabilize soil acidity and prevent phosphorus leaching increasing phosphorus status of the soils [22].

Table 7 shows post-harvest soil levels of potassium. Post harvestings of experimental plants with various methods using *Spirulina* have reported with increased levels of potassium. High amounts of potassium were found in soils of tomato plants with experimental method *Spirulina* + biofertilizer in 25: 75 proportions. The high values were followed by green gram plants in *Spirulina* + vermicompost (25:75) method and 3 h time period soaking. Soil health reflects the soil mineral status.

Increased phosphorous and Potassium levels were observed in both seasons when mung bean plants are supplemented with phosphorus and potassium fertilization [25].

Sustainable agricultural practices increase the soil health and fertility. The sustainable agricultural practices include bio inoculation of soils with marine microalgae, cyanobacteria and combination methods including bio fertilizers, green manures and vermin compost. These methods improve overall soil health and crop yields. These methods also ensure the environmental safety and food security [16].

**Table 7** Effect of *Spirulina* supplementation on % potassium in post-harvest soil of experimental plants

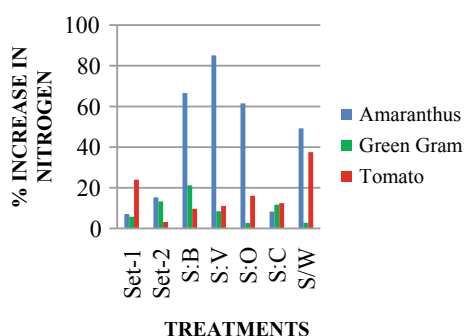
Experimental methods	Potassium (mg/g)		
	A	G	T
Time period soaking	77.5 ± 0.70 (O.N)	150 ± 0.70 (3 h)	184 ± 2.10 (3 h)
Soaking in different concentration	86.5 ± 0.70 (20 g)	78 ± 1.41 (20 g)	77.5 ± 0.7 (5 g)
S + Biofertilizer	89.0 ± 0.70 (in control)	208 ± 1.41 (75:25)	180 ± 0.7 (25:75)
S + Vermicompost	103 ± 0.70 (25:75)	166 ± 4.24 (25:75)	128 ± 0.7 (75:25)
S + Organic manure	76.0 ± 2.12 (in control)	125 ± 2.12 (in control)	153 ± 4.2 (in control)
S + Chemical fertilizer	82.5 ± 0.70 (25:75)	94 ± 1.41 (25:75)	98 ± 0.7 (25:75)
Spray method	65.5 ± 0.70 (50 g/5L)	103 ± 0.70 (100 g/5L)	64 ± 0.70 (50 g/5L)

A Amaranthus; G Green gram, T Tomato

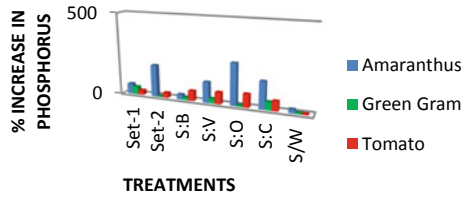
### 3.2 Figures

The Fig. 1 helps to detect the increase in % nitrogen in *Amaranthus*, Green gram and Tomato plants as a comparison. High % of nitrogen is observed for *Amaranthus* among 3 plants and was found in experimental method *Spirulina* + vermicompost.

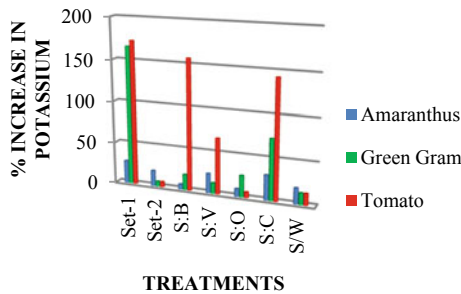
The Fig. 2 describes the % increase of phosphorus compared among three experimental plants. High amounts of phosphorus among 3 plants were observed in *Amaranthus* in *Spirulina* + organic manure method. Figure 3 is about effect of *Spirulina*



**Fig. 1** Percent increase in nitrogen of *Amaranthus*, green gram and tomato by supplementation of *Spirulina* when compared with control \*Set-1 Time period soaking, Set-2 Soaking in different concentrations of *Spirulina*, S:B—*Spirulina*: Biofertilizer, S:V—*Spirulina*: Vermicompost, S:O—*Spirulina*: Organic manure, S:C—*Spirulina*: Chemical Fertilizer, S/W—*Spirulina*/Water



**Fig. 2** Percent increase in phosphorus of *Amaranthus*, green gram and tomato by supplementation of *Spirulina* when compared with control \*Set-1 Time period soaking, Set-2 Soaking in different concentrations of *Spirulina*, S:B—*Spirulina*: Biofertilizer, S:V—*Spirulina*: Vermicompost, S:O—*Spirulina*: Organic manure, S:C—*Spirulina*: Chemical Fertilizer, S/W—*Spirulina*/Water



**Fig. 3** Effect of *Spirulina* supplementation on potassium of *Amaranthus*, green gram and tomato yield when compared with control \*Set-1 Time period soaking, Set-2 Soaking in different concentrations of *Spirulina*, S:B—*Spirulina*: Biofertilizer, S:V—*Spirulina*: Vermicompost, S:O—*Spirulina*: Organic manure, S:C—*Spirulina*: Chemical Fertilizer, S/W—*Spirulina*/Water

on 3 experimental plants in increasing the potassium levels by different experiment methods. High % of potassium was found in tomato plants in 5 h soaking method.

The growth periods of three types of plants used in the experiment were different. Here, only the influence on nutrient status in pre and post experimental treatments in crops and soil were discussed due to difference in growth, morphology, fruit and time of harvest and yields. This can be evident by evaluation of NPK in 3 plants as given in Figs. 1, 2 and 3. The experimental crops were compared with controls.

### 4 Conclusion

The overall outcome of the experiment evaluated was the nutrient profile of crops and postharvest soil with the use of *Spirulina* and other fertilizer combinations and their role in increasing the mineral status. The NPK had shown increase in overall growth of morphological parameters of plants and also maintained the soil fertility. The total increase in growth of plants also results in increase of nutrient content in the plants. In the field study the supplementation of *Spirulina* through different methods

and also in combination with other biofertilizers, vermicompost, organics manure and chemical fertilizers had been used. The synergistic effect of these combinations had shown significant impact on the growth of plants. The soil supplementation of *Spirulina platensis* to plants had improved the soil mineral content. The increased soil phosphorus and potassium had greatly influenced the availability of iron and zinc to the plants. The soil nitrogen influenced the protein content in the plants and also the concentration of pigments like chlorophyll has been increased.

The NPK levels in the soils were increased due to the supplementation of *Spirulina* to the soil and in turn resulted in the overall growth and enhanced the minerals levels in the crops. The potassium is much required for water movement and stomatal movement in the leaves. Thus increased potassium levels helped the crops to uptake sufficient waters and also zinc by foliar application results in good overall growth of plants. This was evident by previous studies in which *Spirulina* application has increased the zinc levels in plants [17].

With the observed results, it can be concluded that *Spirulina platensis* is not only a simple protein supplement, but it can be a good biofortification agent to crops to enhance their mineral nutrient status. Further, this can be proved by complete bioavailability studies and molecular studies. Finally, the present study proves the efficiency of *Spirulina platensis* as a biofortification agent as it has enhanced the mineral nutrient level in the yield of all the three crops and also improves the soil fertility.

## References

1. A. Morte, A. Gutiérrez, B. Dreyer, P.Y. Torrente, M. Honrubia, Biofertilizantes de última generación. Retrieved Apr 27 (2008)
2. A.A. Ibiene, J.U. Agogbua, I.O. Okonko, G.N. Nwachi, Plant growth promoting rhizobacteria (PGPR) as biofertilizer: Effect on growth of *Lycopersicon esculentum*. *J. Am. Sci.* **8**(2), 318–324 (2012)
3. K.D. Hirschi, Nutrient biofortification of food crops. *Annu. Rev. Nutr.* **29**, 401–421 (2009)
4. J. Alexander, J.V. Stein, M. MatinQaim, P. Nestel, H.P.S. Sachdev, Z.A. Bhutta, Analyzing the health benefits of biofortified staple crops by means of the disability-adjusted life years. Approach: a Handbook Focusing on Iron, Zinc and Vitamin A, *International Food Policy Research Institute* (2005)
5. M. Schuler, P. Bauer, *Strategies for Iron Biofortification of Crop Plants, Food Quality*, by K. Kipiris (Ed.), InTech (2012). ISBN: 978-953-51-0560-2
6. M. Ravi, S. De Lata, S. Azharuddin, S.F.D. Paul, The beneficial effects of *Spirulina* focusing on its immunomodulatory and antioxidant properties. *Nutr. Dietary Suppl.* **2**, 73–83 (2010)
7. C.H. Fiske, Y. Subbarow, The colorimetric determination of phosphorus. *J. Biol. Chem.* **66**, 375–400 (1925)
8. G.H. Jeffery, J. Bassett, J. Mendham, R.C. Denney, Vogel's textbook of quantitative chemical analysis. 5th. ed. Harlow: Longman Scientific & Technical, 906 (1989)
9. A. Boraste, K.K. Vamsi, A. Jhadav, Y. Khairnar, N. Gupta, S. Trivedi, P. Patil, G. Gupta, M. Gupta, A.K. Mujapara, B. Joshi, Biofertilizer: A novel tool for agriculture. *Int. J. Microbiol. Res.* **1**(2), 23 (2009)

10. E. Kisetu, Z.S. Mtakimwa, Incorporating pigeon pea compost with Minjingu fertilizer brands to determine their effects on maize production in Morogoro Tanzania. *World J. Agric Sci.* **1**(9), 294–298 (2013)
11. S.T. Zodape, S. Mukhopadhyay, K. Eswaran, M.P. Reddy, J. Chikara, Enhanced yield and nutritional in green gram (*Phaseolus radiata* L) treated with seaweed (*Kappaphycus alvarezii*) extract. *J. Sci. Ind. Res.* **69**, 468–471 (2010)
12. F.C. Gómez-Merino, L.I. Trejo-Téllez, Biostimulant activity of phosphite in horticulture. *Sci. Hortic.* **196**, pp. 82–90 (2015). <https://doi.org/10.1016/j.scienta.2015.09.035>. (http://www.sciencedirect.com/science/article/pii/S0304423815301990). ISSN: 0304-4238
13. R.N. Padhy, N. Nayak, R.R. Dash-Mohini, S. Rath, R.K. Sahu, Growth, metabolism and yield of rice cultivated in soils amended with fly Ash and cyanobacteria and metal loads in plant parts, *Rice Sci.* **23**(1), 22–32 2016, ISSN: 1672-6308. <https://doi.org/10.1016/j.rsci.2016.01.003>
14. N. Bidyarani, R. Prasanna, S. Babu, F. Hossain, A.K. Saxena, Enhancement of plant growth and yields in Chickpea (*Cicerarietinum* L.) through novel cyanobacterial and biofilmed inoculants. *Microbiol. Res.* **188–189**, 97–105 (2016), ISSN: 0944-5013. <https://doi.org/10.1016/j.micres.2016.04.005>
15. S. Araghian, A. Bagherzadeh, R. Haghighi, Effect of brown algae and vermicompost application on some cherry tomato traits in hydroponic system. **10**, 77–83 (2015)
16. S.M. Hamed, A.A.A. El-Rhman, N. Abdel-Raouf, B.M.I. Ibraheem, Role of marine macroalgae in plant protection & improvement for sustainable agriculture technology, Beni-Suef Univ. J. Basic Appl. Sci. **7**(1) 2018, 104–110, ISSN: 2314-8535. <https://doi.org/10.1016/j.bjbas.2017.08.002>
17. L. Anitha, G. SaiBramari, P. Kalpana, Effect of supplementation of *Spirulina platensis* to enhance the Zinc status in plants of *Amaranthus gangeticus*, *Phaseolusaureus* and tomato. *Adv. Biosci. Biotechnol.* **7**, 289–299 (2016)
18. S.F. Bender, C. Wagg, M.G.A. Van Der Heijden, An underground revolution: biodiversity and soil ecological engineering for agricultural sustainability. *Trends Ecol. Evol.* **31**, 440–452 (2016). <https://doi.org/10.1016/j.tree.2016.02.016>
19. R.S. Sharifi, Study of nitrogen rates effects and seed biopriming with PGPR on quantitative and qualitative yield of Safflower (*Carthamus tinctorius* L.) *Tech. J. Eng. Appl. Sci.* **2**, 162–166 (2012)
20. B. Singh, K. Singh, D. Talwar, S.K. Jindal, V.S. Sardana, Influence of bio-fertilizers on growth and yield attributing attributes in Tomato. *Int. J. Curr. Microbiol. App. Sci.* **7**(4), 3686–3694 (2018). <https://doi.org/10.20546/ijcmas.2018.704.414>
21. J.S. Singh, A. Kumar, A.N. Rai, D.P. Singh, Cyanobacteria: A precious bio-resource in agriculture, ecosystem, and environmental sustainability. *Frontiers Microbiol* **7**, 529 (2016). <https://doi.org/10.3389/fmicb.2016.00529>
22. G.A. Dias, R. Rocha, J.L. Araújo, J.F. Lima, W.A. Guedes, Growth, yield, and postharvest quality in eggplant produced under different foliar fertilizer (*Spirulina platensis*) treatments. *Semina: Ciências Agrárias.* **37**. 3893 (2016). <https://doi.org/10.5433/1679-0359.2016v37n6p3893>
23. A.J. Stein, J.V. Meenakshi, M. Qaim, P. Nestel, H.P.S. Sachdev, Z.A. Bhutta, Analyzing the health benefits of biofortified staple crops by means of the disability-adjusted life years approach: a handbook focusing on Iron, Zinc and Vitamin A. *Int. Food Policy Res. Inst.* (2005)
24. A.Z. Hegazi, S.S.M. Mostafa, H.M.I. Ahmed, Influence of different cyanobacterial application methods on growth and seed production of common bean under various levels of mineral nitrogen fertilization. *Nat. Sci.* **8**(11), 183–194 (2010)
25. M.M. Ibrahim, M.S.S. Al- Bassyuni, Effect of irrigation intervals, phosphorus and potassium fertilization rates on productivity and chemical constituents of mung bean plants. *Res. J. Agric. Biol. Sci.* **8**(2): 298–304, 2012, ISSN 1816-1561

# GC-MS Analysis and Computational Studies of Roots of *Anthocephalus Cadamba*



Kaveripakam Sai Sruthi and Adikay Sreedevi

**Abstract** The current investigation was aimed at identification of various bioactive compounds present in ethanol extract of *Anthocephalus cadamba* roots using GC-MS analysis and to perform molecular docking analysis. The ethanol extract of roots was prepared using hot extraction method and submitted to GC-MS analysis. Biological activities were predicted for all separated components using Dr. Duke's ethnobotanical database. The identified bioactive compounds which were predicted to possess antioxidant properties were subjected to computational studies. Primarily drug like properties and bioactivity score was calculated and then the compounds were subjected to docking studies using GOLD software to ascertain the binding fitness score towards Bcl2 and NADPHoxidase. The outcome of GC-MS analysis reveals that roots of *Anthocephalus cadamba* are rich source of bioactive phytoconstituents which may ladle out as leads in development of drugs that possess good antioxidant properties and thereby aids in treatment of various related disorders like nephrotoxicity, cancers, diabetes and obesity.

**Keywords** *Anthocephalus cadamba* · Docking · Anti-oxidant

## 1 Introduction

Plants have established the knowledgeable conventional medicine systems that were in existence from the time immemorial [1]. Medicinal plants are encountered with outstanding history in several indigenous communities and continuously providing useful strategies for treating various ailments [2]. In the recent years there was raising interest in identification of phytoconstituents and exploring the pharmacological activities of various traditional medicinal herbs, in view of the natural origin, cost effectiveness, and fewer side effects. It has been evidenced that phytoconstituents

---

K. S. Sruthi (✉) · A. Sreedevi

Division of Pharmaceutical Chemistry, Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India  
e-mail: [sruthisai7@gmail.com](mailto:sruthisai7@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_50](https://doi.org/10.1007/978-3-030-46939-9_50)

569



either as pure compounds or as standardized extracts of medicinal herbs provide numerous opportunities for novel drug leads.

Among many medicinal plants used in traditional medicine, *Anthocephalus cadamba* belonging to the family Rubiaceae is an ayurvedic remedy in the treatment of several ailments [3]. But till today the plant is not well scientifically exploited to identify the various bioactive phytoconstituents present in it. Hence the current investigation was aimed at identification of various phytoconstituents in roots of this plant by employing GC-MS technique and further to perform computer aided molecular docking of major phytoconstituents to evaluate their imperative biological role as antioxidants.

## 2 Methodology

### 2.1 Collection and Authentication of Plant Materials

Roots of *Anthocephalus cadamba* were self possessed from Tirumala hills in chittoor district. Roots were authenticated by Botanist and specimen (No: 1856) was put in Botany department, Sri Venkateswara University, Tirupati, India.

### 2.2 Preparation of Extract

Collected roots were dried in shade and powdered. Root powder was defatted with petroleum ether and the obtained marc was then macerated with ethanol for 24 h, refluxed for 3 h and filtered. The procedure was repeated thrice and the filtrates were combined and distilled. The ethanol extract thus obtained was air dried and allowed to concentrate under reduced pressure to get the semi solid residue.

### 2.3 Gas Chromatography Associated with Mass Spectrometry

The prepared ethanol extract of *Anthocephalus cadamba* (EEAC) was subjected for GC-MS analysis which was carried out in a GC-MS system (Clarus 500 Perkin-Elmer with turbomass spectrometer (5.2.0)) The fused silica capillary column (30 m; 500  $\mu$ m id) was used. Injection port 280°C; flow rate is 1.0 ml/min; ionization voltage 70 eV. The sample injected in split mode at 1:10 and the obtained peaks in GC were subjected to Mass Spectroscopy. Mass spectral scan range at 40–450 amu; ion source temperature 160 °C; Interface temperature 180 °C and the end time 54.5 min.

**Table 1** Selected ligands

Compound code	Ligand name
L-1	1(2H)-Naphthalenone, 3,4,5,6,7,8-hexahydro-
L-2	2,4-Dihydroxy-2,5-dimethyl-3(2H)-furan-3-one
L-3	2,5-Dimethyl-4-hydroxy-3(2H)-furanone
L-4	4-((1E)-3-Hydroxy-1-propenyl)-2-methoxyphenol
L-5	4H-Pyran-4-one, 2,3-dihydro-3,5-dihydroxy-6-methyl-
L-6	9-Octadecene, (E)-
L-7	Decahydro-2,2-dimethyl-naphthalene
L-8	Ethyl $\alpha$ -d-glucopyranoside
L-9	n-Hexadecanoic acid

Interpretation of GC-MS was executed by utilizing the National Institute of Standard and Technology's database. Further bioactivity of the identified components was predicted using Dr. Duke's ethnomedicinal database.

## 2.4 Computational Studies

Based on their predicted bioactivity in ethno-medicinal database among all the compounds, 9 phytoconstituents were selected for computational studies (Table 1).

## 2.5 Prediction of Molecular Properties

Lipinski's rule of five was employed to determine drug likeness and also to estimate whether a chemical substance predicted or possessing with some biological activity consists of properties to be orally active [4]. By using Molinspiration, an online server calculation of essential molecular properties like "molecular weight, log P, hydrogen bond acceptor and donor of selected ligands" for selected ligands was done.

## 2.6 Calculation of Bioactivity Score

Bioactivity of different selected ligands was determined by assessing the activity score of "GPCR ligand, ion channel inhibitor, Kinase inhibitor, nuclear receptor ligand, protease inhibitor, enzyme inhibitor" using molinspiration server.

## 2.7 Docking

The Bcl-2 (PDB ID: 1G5M) and NADPH oxidase (PDB ID: 3A1F) structures were acquired from PDB database. Then adopting SPDBV software the chains which are unnecessary and hetero atoms were removed. The recognition of active site in respective proteins was done by CASTp server. Docking of compounds with respective proteins was performed using GOLD software (3.0.1 version) [5].

Gold Score carries a force field based scoring function and the total fitness score is computed:

$$\text{GoldScore} = S(\text{hb\_ext}) + S(\text{vdw\_ext}) + S(\text{hb\_int}) + S(\text{vdw\_int})$$

Where

S (hb\_ext) is protein-ligand hydrogen bond score,

S (vdw\_ext) is protein-ligand van der Waals score,

S (hb\_int) is the score from intramolecular hydrogen bond in the ligand and

S (vdw\_int) is the score from intramolecular strain in the ligand.

## 3 Results

### 3.1 GC-MS Studies of Ethanol Extract of Anthocephalus Cadamba

GC-MS studies of EEAC revealed the existence of 39 chemical constituents. The Retention time ( $R_t$ ), Molecular formula, Molecular weight and concentration of identified compounds are illustrated in Table 2. EEAC's GC-MS chromatogram was depicted in Fig. 1.

### 3.2 Molecular Properties

The calculated molecular properties of the phytoconstituents were tabulated in Table 3. Most of the selected compounds have number of violations equal to zero. All the selected phytoconstituents obeyed Lipinski's rule and showed drug like property.

**Table 2** List of compounds identified in EEAC by GC-MS analysis

S. No.	Peak name	Retention time	Peak area	% Peak area	Bioactivity <sup>a</sup>
1	Name: 1-Butanol, 3-methyl- Formula: C <sub>5</sub> H <sub>12</sub> O; MW: 88	2.76	110,706,040	24.0861	COMT inhibitor
2	Name: 3-Amino-2-oxazolidinone Formula: C <sub>3</sub> H <sub>6</sub> N <sub>2</sub> O <sub>2</sub> ; MW: 102	3.51	4,297,269	0.9349	Aromatic amino acid decarboxylase activity enhancer
3	Name: 2-Furanmethanol Formula: C <sub>5</sub> H <sub>6</sub> O <sub>2</sub> ; MW: 98	4.84	1,786,246	0.3886	Nf
4	Name: 2-Hydroxy-3-pentanone Formula: C <sub>5</sub> H <sub>10</sub> O <sub>2</sub> ; MW: 102	5.26	419,617	0.0913	Aryl hydrocarbon hydroxylase inhibitor
5	Name: Pyrazine, 2,5-dimethyl- Formula: C <sub>6</sub> H <sub>8</sub> N <sub>2</sub> ; MW: 108	5.60	2,199,101	0.4785	Nf
6	Name: Furan, tetrahydro-3-methyl- Formula: C <sub>5</sub> H <sub>10</sub> O; MW: 86	5.87	2,016,432	0.4387	COMT inhibitor
7	Name: 1,2-Cyclopentanedione Formula: C <sub>5</sub> H <sub>6</sub> O <sub>2</sub> ; MW: 98	6.25	13,773,966	2.9968	Nf
8	Name: 2,4-Dihydroxy-2,5-dimethyl-3(2H)-furan-3-one Formula: C <sub>6</sub> H <sub>8</sub> O <sub>4</sub> ; MW: 144	7.04	3,436,779	0.7477	HDL-genic, hepatoprotective
9	Name: Pyrazine, trimethyl- Formula: C <sub>7</sub> H <sub>10</sub> N <sub>2</sub> ; MW: 122	7.53	1,175,766	0.2558	Nf
10	Name: Dimethylamine, N-(neopentyl-oxyl)- Formula: C <sub>7</sub> H <sub>17</sub> NO; MW: 131	8.56	28,536,738	6.2087	Anti-tumor, NADH-oxidase inhibitor, NO scavenger
11	Name: 2,5-Dimethyl-4-hydroxy-3(2H)-furanone Formula: C <sub>6</sub> H <sub>8</sub> O <sub>3</sub> ; MW: 128	9.55	5,883,554	1.2801	HDL-genic, Hepatotoxic, hypolipidemic

(continued)

Table 2 (continued)

S. No.	Peak name	Retention time	Peak area	% Peak area	Bioactivity <sup>a</sup>
12	Name: Ethanone, 1-(1-cyclohexen-1-yl)- Formula: C <sub>8</sub> H <sub>12</sub> O; MW: 124	9.95	9,172,419	1.9956	Nf
13	Name: Spiro[tetrahydrofuran-3,5'-hydantoin] Formula: C <sub>6</sub> H <sub>8</sub> N <sub>2</sub> O <sub>3</sub> ; MW: 156	10.43	8,947,070	1.9466	Nf
14	Name: 4H-Pyran-4-one, 2,3-dihydro-3,5-dihydroxy-6-methyl- Formula: C <sub>6</sub> H <sub>8</sub> O <sub>4</sub> ; MW: 144	11.71	13,566,533	2.9516	Anti oxidant, anti-inflammatory, hypolipidemic
15	Name: 2-Propyl-tetrahydropyran-3-ol Formula: C <sub>8</sub> H <sub>16</sub> O <sub>2</sub> ; MW: 144	11.86	4,773,552	1.0386	Nf
16	Name: Octanoic Acid Formula: C <sub>8</sub> H <sub>16</sub> O <sub>2</sub> ; MW: 144	12.28	1,177,506	0.2562	Aromatic amino acid decarboxylase activity enhancer
17	Name: Benzenecarboxylic acid Formula: C <sub>7</sub> H <sub>6</sub> O <sub>2</sub> ; MW: 122	13.04	9,296,275	2.0226	Arachadonic acid inhibitor
18	Name: Phenol, 4-ethyl-2-methoxy- Formula: C <sub>9</sub> H <sub>12</sub> O <sub>2</sub> ; MW: 152	14.33	6,987,214	1.5202	Nf
19	Name: 2-Methoxy-4-vinylphenol Formula: C <sub>9</sub> H <sub>10</sub> O <sub>2</sub> ; MW: 150	15.40	11,844,731	2.5770	Nf
20	Name: 1(2H)-Naphthalenone, 3,4,5,6,7,8-hexahydro- Formula: C <sub>10</sub> H <sub>14</sub> O; MW: 150	15.63	11,158,841	2.4278	HDL-genic, HMG-Co-A activity inhibitor
21	Name: Phenol, 2-methoxy-4-(1-propenyl)- Formula: C <sub>10</sub> H <sub>12</sub> O <sub>2</sub> ; MW: 164	16.28	6,645,700	1.4459	Nf
22	Name: Decanoic acid, ethyl ester Formula: C <sub>12</sub> H <sub>24</sub> O <sub>2</sub> ; MW: 200	16.62	19,244,438	4.1870	Arachidonic acid inhibitor

(continued)

Table 2 (continued)

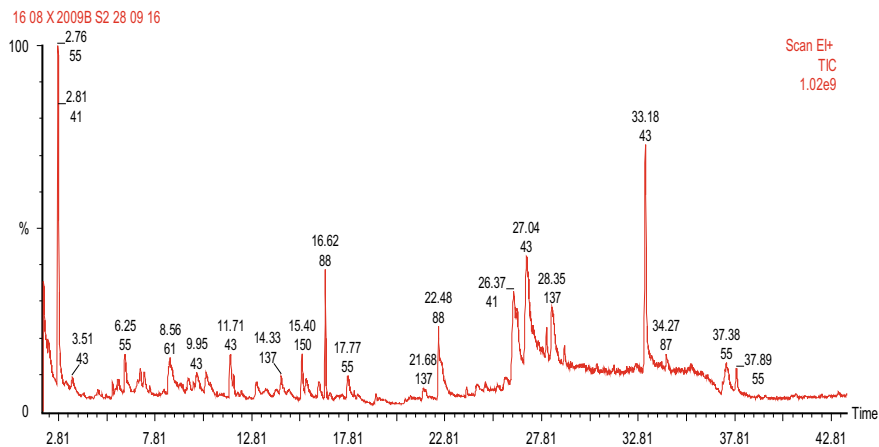
S. No.	Peak name	Retention time	Peak area	% Peak area	Bioactivity <sup>a</sup>
23	Name: 2,4-Pentadien-1-ol, 3-ethyl-, (2Z)- Formula: C <sub>7</sub> H <sub>12</sub> O; MW: 112	17.77	10,749,416	2.3387	Zinc bioavailability enhancer
24	Name: 3,3-Dimethyl-4-heptanol Formula: C <sub>9</sub> H <sub>20</sub> O; MW: 144	18.11	685,042	0.1490	Nf
25	Name: Phenol, 2-methoxy-4-(1-propenyl)- Formula: C <sub>10</sub> H <sub>12</sub> O <sub>2</sub> ; MW: 164	19.23	2,368,582	0.5153	Antioxidant
26	Name: Naphthalene, decahydro-2,2-dimethyl- Formula: C <sub>12</sub> H <sub>22</sub> ; MW: 166	20.77	557,999	0.1214	HMG co.A reductase inhibitor, hypocholesterolemic
27	Name: 2-Propanone, 1-(4-hydroxy-3-methoxyphenyl)- Formula: C <sub>10</sub> H <sub>12</sub> O <sub>3</sub> ; MW: 180	21.68	3,129,303	0.6808	Aryl hydrocarbon hydroxylase inhibitor
28	Name: Dodecanoic acid, ethyl ester Formula: C <sub>14</sub> H <sub>28</sub> O <sub>2</sub> ; MW: 228	22.48	6,894,895	1.5001	Arachidonic acid inhibitor
29	Name: 1H-Pyrazole-1-carboxaldehyde, 4-ethyl-4,5-dihydro-5-propyl- Formula: C <sub>9</sub> H <sub>16</sub> N <sub>2</sub> O; MW: 168	24.47	5,515,264	1.1999	Hematinic, hypotensive
30	Name: Phenol, 3,4,5-trimethoxy- Formula: C <sub>9</sub> H <sub>12</sub> O <sub>4</sub> ; MW: 184	24.88	3,494,323	0.7603	Nf
31	Name: 1,4-Undecadiene, (Z)- Formula: C <sub>11</sub> H <sub>20</sub> ; MW: 152	25.54	1,971,145	0.4289	Zinc bioavailability enhancer
32	Name: Ethyl ß-D-glucopyranoside Formula: C <sub>8</sub> H <sub>16</sub> O <sub>6</sub> ; MW: 208	25.93	6,070,897	1.3208	Anti-cancer, SOD enhancer, diuretic

(continued)

Table 2 (continued)

S. No.	Peak name	Retention time	Peak area	% Peak area	Bioactivity <sup>a</sup>
33	Name: 2H-Pyran-2-one, 5-ethyridenetetrahydro-4-(2-hydroxyethyl)- Formula: C <sub>9</sub> H <sub>14</sub> O <sub>3</sub> ; MW: 170	26.37	20,297,012	4.4160	Nf
34	Name: Tetradecanoic acid Formula: C <sub>14</sub> H <sub>28</sub> O <sub>2</sub> ; MW: 228	28.06	5,451,068	1.1860	Arachidonic acid inhibitor
35	Name: 4-((1E)-3-Hydroxy-1-propenyl)-2-methoxyphenol Formula: C <sub>10</sub> H <sub>12</sub> O <sub>3</sub> ; MW: 180	28.35	28,711,382	6.2467	Anti-cancer, Trypsin-enhancer, Fibrinolysis enzyme activator
36	Name: n-Hexadecanoic acid Formula: C <sub>16</sub> H <sub>32</sub> O <sub>2</sub> ; MW: 256	33.18	64,897,784	14.1197	Anti-tumor, TNF- $\alpha$ production inhibitor, NO scavenger
37	Name: 4-Oxo- $\delta$ -isodamascol Formula: C <sub>13</sub> H <sub>20</sub> O <sub>2</sub> ; MW: 208	34.27	3,776,570	0.8217	Nf
38	Name: 9-Octadecene, (E)- Formula: C <sub>18</sub> H <sub>36</sub> ; MW: 252	37.30	12,023,304	2.6159	Anti-cancer, fibrinolysis enzyme activator
39	Name: Octadecanoic acid Formula: C <sub>18</sub> H <sub>36</sub> O <sub>2</sub> ; MW: 284	37.90	5,985,907	1.3023	Arachidonic acid inhibitor

<sup>a</sup>Source Dr. Duke's ethnomedicinal database



**Fig. 1** GC-MS chromatogram of ethanol extract of *Anthocephalus cadamba*

**Table 3** Calculated molecular properties of phytoconstituents

Compound	miLogP	TPSA	MW	nON	nOHNH	nvio	nrotb	vol
L-1	2.20	17.07	150.22	1	0	0	0	154.96
L-2	-0.02	66.76	144.13	4	2	0	0	122.84
L-3	0.55	46.53	128.13	3	1	0	0	115.15
L-4	1.37	49.69	180.20	3	2	0	3	169.84
L-5	-0.20	66.76	144.13	4	2	0	0	123.40
L-6	8.80	0.00	252.49	0	0	1	14	308.40
L-7	4.04	0.00	156.23	0	0	0	1	161.40
L-8	-1.65	99.38	208.21	6	4	0	3	186.14
L-9	7.06	37.30	256.43	2	1	1	14	291.42

Log *p*—partition coefficient nrotb—No of rotatable bonds TPSA—Topological polar surface area  
nvio—No of violations

nON—No of hydrogen bond acceptors MW—Molecular weight nOHNH—No of hydrogen bond donars Vol—volume

### 3.3 Bioactivity Score Prediction

The computed bioactivity scores of the screened compounds for G protein coupled receptor, Ion channel inhibitor, Kinase inhibitor, Nuclease receptor, protease inhibitor and enzyme inhibitor were summarized in Table 4.



**Table 4** Predicted bioactivity score of phytoconstituents

Compound	GPCR ligand	Ion channel inhibitor	Kinase inhibitor	Nuclear receptor ligand	Protease inhibitor	Enzyme inhibitor
L-1	-1.18	-0.51	-2.01	-0.82	-1.19	-0.41
L-2	-1.60	-0.64	-1.98	-1.29	-1.54	-0.52
L-3	-2.73	-1.80	-3.43	-2.57	-2.78	-1.61
L-4	-0.55	-0.05	-0.74	-0.30	-1.00	-0.08
L-5	-1.29	-0.77	-2.08	-1.53	-1.14	-0.08
L-6	-0.08	0.02	-0.28	-0.11	-0.23	0.07
L-7	-0.70	-0.22	-0.92	-0.79	-0.86	-0.34
L-8	-0.25	0.10	-0.54	-0.78	-0.40	0.38
L-9	0.02	0.06	-0.33	0.08	-0.04	0.18

**Table 5** Fitness score of selected phytoconstituents to Bcl-2

Compound	S(hb_ext)	S(vdw_ext)	S(hb_int)	S(int)	Fitness
L-1	0.00	14.06	0.00	0.00	19.34
L-2	9.81	5.86	0.00	-2.05	15.83
L-3	9.82	1.89	0.00	0.00	12.42
L-4	4.41	14.70	0.00	-6.80	17.81
L-5	0.00	12.21	0.00	-4.72	12.07
L-6	0.00	24.48	0.00	-8.15	25.52
L-7	0.00	13.77	0.00	-3.21	15.72
L-8	3.99	11.88	0.00	-6.31	14.02
L-9	5.80	22.12	0.00	-11.62	24.59

### 3.4 Docking Studies

The results of docking studies were assessed based on binding compatibility (Docked energy-kcal/mol). The GOLD docking scores for the compounds were illustrated in Table 5 and 6.

## 4 Discussion

GC-MS analysis is one of the vital tools for the reliable identification of phytoconstituents present in plant extracts [6]. It is an efficient way to analyze the fingerprinting of phytomedicine and to evaluate the overall chemical difference in medicinal plants that provides high separation efficiencies [7, 8]. The current study was aimed to

**Table 6** Fitness score of selected phytoconstituents to NADPH oxidase

Compound	S(hb_ext)	S(vdw_ext)	S(hb_int)	S(int)	Fitness
L-1	0.00	22.30	0.00	0.00	30.67
L-2	2.00	16.58	0.00	-0.95	23.85
L-3	16.02	6.98	0.00	0.00	25.62
L-4	0.46	24.48	0.00	-7.11	27.01
L-5	2.00	17.98	0.00	-4.87	21.86
L-6	0.00	28.05	0.00	-8.32	30.25
L-7	0.00	23.64	0.00	-3.21	29.29
L-8	1.83	20.53	0.00	-6.62	23.43
L-9	0.00	26.55	0.00	-7.44	29.07

find out the bioactive compounds existed in the EEAC by using gas chromatography and mass spectroscopy. The results pertaining to the GC-MS analysis led to the identification of 39 compounds in EEAC. The gas chromatogram shows the relative concentration of the components present in the plants. The heights of the peaks indicate the relative concentrations of the components present in the plant. These reports are the first of their kind to analyze the chemical constituents of roots of *Anthocephalus cadamba*. Among the identified chemical constituents, some of them are predicted to possess antioxidant properties which in turn aids in the cure of disorders like obesity, hepatotoxicity, nephrotoxicity and cancers.

In this context, to support their predicted bioactivity in silico studies was carried out. The utilization of these computational screening techniques helps to understand pharmacological behavior and reduce the cost of random preclinical screening and further accelerates the process [9]. Based on predicted bioactivity in ethnomedicinal database some of the phytoconstituents identified in GC-MS EEAC were selected for molecular docking studies.

Initially the molecular descriptors of the selected phytoconstituents were determined to assess the drug likeness. "The application of drug likeness has gained wide acceptance as an approach to reduce attrition in drug discovery and development" [10]. Lipinski's rule of five is widely accepted to predict the molecular properties that are prominent for pharmacokinetics of drugs in vivo [4]. All the compounds (as number of violations  $\leq 1$ ) were found to obey Lipinski's rule of five.

The bioactivity of the compounds were predicted and the bioactive score of these compounds indicated good to moderate activity towards GPCR, Ion channel inhibitors, Kinase inhibitor, Nuclear receptor ligands, Protease inhibitor and Enzyme inhibitor. These scores for organic molecules can be represented as active (bioactive score  $>0$ ), moderately active (bioactive score 0 to  $-5$ ) and inactive (bioactive score  $>-5$ ) [11, 12].

Gold software was used to identify the best ligand to interact with selected receptors/proteins namely Bcl-2 and NADPH oxidase. Bcl-2 regulates cell death and is considered as an anti-apoptotic protein [13]. Apoptosis is the major condition seen in

pathology of many disorders where antioxidants levels are declined and free radical scavenging activity is reduced. The binding interactions of the ligands with this Bcl-2 protein was evaluated and the highest binding fitness with Bcl-2 was found to be for 9-Octadecene, (E)-(L-6) followed by n-Hexadecanoic acid (L-9) which may be due to the strong binding interactions of ligands at the different active site residues or due to vander waals energy or may be because of ligand torsion strain [14].

NADPH (Nicotinamide adenine dinucleotide phosphate) oxidases are multisubunit enzymes that transfer electrons across cell membranes. Research on underlying pathogenesis of many diseases suggested an important role of oxidative stress. Among numerous internal sources of ROS NADPH oxidase is generally accepted as major producer [15]. In case of NADPH oxidase, all the phytoconstituents exhibited good binding fitness. The highest binding fitness with NADPH oxidase was found to be for 1(2H)-Naphthalenone, 3, 4, 5, 6, 7, 8-hexahydro- (L-1) followed by 9-Octadecene, (E)- (L-6) which may be due to strong binding interactions with active site amino acid residues or due to vanderwaals energy [16].

## 5 Conclusion

The findings of study reveals that the roots of *Anthocephalus cadamba* are rich source of potent bioactive phytoconstituents which may ladle out as leads in the development of drugs that possess good antioxidant properties and thereby aids in treatment of various related disorders like nephrotoxicity, cancers, diabetes and obesity.

## References

1. R. Nishi, M. Radha, Pharmacognostic and physicochemical analysis on the leaves of *Brufeslsia americana* L. Asian Pac. J. Trop. Biomed. s305–307 (2012)
2. K. Bairwa, R. Kumar, R.J. Sharma, R.K. Roy, An updated review on *Bidons Pilosa l.* Der. Pharma. Chemical. **2**, 325–337 (2010)
3. K.R., Kirtikar, B.D., Basu, *Indian medicinal plants*, 2nd edn, Vailey offset publishers Dehra dun, pp. 1998 (18480)
4. C.A., Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug. Delivery. Rev. **23**, 4–25 (1997)
5. M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R.D. Taylor, Improved protein–ligand docking using GOLD. Proteins. **52**, 609–623 (2003)
6. J. Narayanan, J. Antonysam, GC-MS analysis of ethanolic extracts of *Cyatheanil girensis*, *C.gigantea*, and *C.crinita*. Egypt. Pharmaceut. J. **15**, 43–47 (2017)
7. R. Hall, M. Beale, O. Fiehn, N. Hardy N, L. Sumner, Plant metabolomics: the missing link in functional genomics strategies. Plant cell. **14**, 1437–1440 (2002)
8. L.W. Sumner, P. Mendes, R.A. Dixon, Plant metabolomics: the mixing link in functional genomics era. Phytochem. **62**, 817–836 (2003)
9. S. Ekins, J. Mestres, B. Testa, *silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br. J. Pharmacol. **152**, 9–20 (2007)

10. S. Chandra, K. Thilak, Molecular properties prediction, docking studies and antimicrobial screening of ornidazole and its derivatives. *J. Chem. Pharm.* **8**, 849–861 (2016)
11. G. Valli, A. Jayalakshmi, Assessment of molecular property and bio-activity scores of imines derived from disubstituted aromatic aldehydes and 4-amino antipyrine using molinspiration. *Int. J. Chem. Con.* **1**, 175–179 (2015)
12. S. Singh, A. Gupta, A. Verma, Molecular properties and bioactivity score of the *Aloe vera* antioxidant compounds—in order to lead finding. *Res. J. Pharm. Biol. Chem. Sci.* **4**, 876–881 (2013)
13. N. Aaron, A. Jeffrey, C. Anthony, The BCL-2 family: key mediators of the apoptotic response to targeted anti-cancer therapeutics. *Cancer Discov.* **5**, 475–487 (2015)
14. A. Mustaq, J. Kaiser, BCL-2 as target for molecular docking of some neoplastic drugs. *Sci. Rep.* **1**, 458 (2012)
15. A. Fortuno, J.G. San, M.U. Moreno, O. Beloqui, J. Diez, G. Zalba, Phagocytic NADPH Oxidase over activity underlies oxidative stress in metabolic syndrome. *Diabetes* **55**, 209–215 (2006)
16. J. Jiang, H. Kang, X. Song, S. Huang, S. Li, J. Xu, A model of interaction between nicotinamide adenine dinucleotide phosphate (NADPH) oxidase and apocynin analogues by docking method. *Int. J. Mol. Sci.* **14**, 807–817 (2013)

# Comparative Omics Based Approach to Identify Putative Immunogenic Proteins of *Trichomonas Foetus*



Geethanjali Karli, Rathnagiri Polava, and Kalarani Varada

**Abstract** Bovine Trichomonosis is the most neglected sexually transmitted disease in cattle. *Trichomonas foetus* is a flagellate protozoan known for causing simple symptoms like, vaginal discharge to as severe as abortion and infertility. Trichomonosis is very difficult to treat. Screening and isolations of infected animals is the only available strategy. Lack of simple, rapid, inexpensive point of care diagnostic kits is leading to rapid spread. All the published Transcriptomics, proteomics studies of *Trichomonas foetus* and comparative genomics based approach from *Trichomonas vaginalis*, a similar human pathogen were screened thoroughly to list set of highly expressed proteins. The FASTA Sequences of the proteins were analyzed by TMMHM and IEDB resource for identification of Putative membrane proteins and antigenic domains respectively. Cysteine protease 8 was found to be consistently expressed extracellular secretory protein and Surface antigen BspA-like Protein and Chlamydia polymorphic membrane protein-like extracellular domain were found to be highly immunogenic membrane proteins.

**Keywords** Transcriptomics · Proteomics · Putative · Immunogen · Bovine · B-cell epitope prediction · TMMHM

## 1 Introduction

India owns the world's largest livestock population with 528 millions of domestic animals. India ranks first for buffalo with 105.3 million population and second in cattle with 199 million. Livestock population is facing severe stress due to infectious disease outbreaks [1].

---

G. Karli · K. Varada (✉)

Department of Biotechnology, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [kala.dandala@gmail.com](mailto:kala.dandala@gmail.com)

R. Polava  
Genomix Biotech Inc, Atlanta, USA

G. Karli  
Department of Biotechnology, Govt. Degree College, Kukatpally, Hyderabad, India

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_51](https://doi.org/10.1007/978-3-030-46939-9_51)

With regard to cattle, only 28% percent of the cattle breeding is with Artificial insemination, remaining Proportion of the cattle are under natural breeding process with huge threat of disease transmission and genetic abnormalities. The Conception rate of cattle is as low as only 35% in India [1–3].

In the recent years, village farming is transforming into organized farms, using cross bred high quality bulls for breeding. But still there is no practice of disease free certification, which is resulting in uncontrolled spreading of Sexually transmitted diseases like Brucellosis, Infectious Bovine Rhinotracheitis (IBR), Bovine Tuberculosis, Bovine Trichomoniasis, Bovine Genital Campylobacteriosis, Foot and Mouth disease (FMD) and Bovine Viral Diarrhoea (BVD) among bovine population in the country. More over same bulls are repeatedly used for natural service [1].

Most of the diseases are zoonotic, posing public health issues. There are no epidemiological reports due to non- availability of simple, rapid and inexpensive point of care diagnostic kits. The burden is expected to be >1 Billion USD.

### ***1.1 Venereal Diseases of Cattle-proposed Control Strategies***

India has major challenges with regard to low productivity of livestock. Most of the cattle are in natural service with unknown genetic and disease potential. Inadequate coverage through artificial insemination, non-availability of elite males for natural service, poor hygienic conditions are the major causes of low conception, high morbidity & mortality [1, 2].

On recommendation of OIE, Govt. of India and the Animal Husbandry Dairying & Fisheries department has setup a few Standard protocols and operating procedures for bovine breeding. Few tests were prescribed and made mandatory to screen for Brucellosis, Bovine Tuberculosis, Para TB, Trichomonosis, Bovine Genital Campylobacteriosis, FMD, Infectious Bovine Rhinotracheitis (IBR) and Bovine Viral Diarrhoea (BVD). Disease diagnostics laboratories and regional centers were developed with experts from National dairy development Board, Indian veterinary research Institute, NIVEDI, other ICAR allied institutes and Veterinary Universities. The main objective was to improve the productivity of livestock through enhanced AI coverage. As per the impact analysis report submitted by NABARD, overall conception rate has increased from 20 to 35% [1, 3].

In India, only since a decade, few point of care diagnostics like ELISA and Lateral flow assay rapid kit are developed and lab screening facility is available for Brucellosis [5, 6], Para Tuberculosis [4]. Bovine Trichomonosis is the most neglected cattle disease, which is responsible for drastic drop in fertility rate in cattle. Hence this work focuses to screen for potential antigenic genes of *Trichomonas foetus*, a protozoan parasite responsible to cause Trichomonosis in cattle.

## 2 Pathophysiology of Trichomonosis

*T. foetus* causes bovine Trichomonosis in cattle. It has a characteristic pyriform/sperm head body. It has three anterior flagella, undulating membrane and one posterior flagellae. Its typical jerky, rolling motion was detected by light and advanced microscopy [7]. Initially only the trophozoite stage was reported, but few reported the pseudocyst form as well [8]. The “Belfast” strain, is predominant in United states, some European countries and African continent [9]. The strain “Brisbane” is prevalent in Australia [10]. Strain “Manley” was rarely reported [11]. Apart from cattle, *T. foetus* infection were reported in cats and pigs [12, 13]. It is also found to be zoonotic from the reports associated with *T. foetus* found in immunocompromised and immunosuppressed individuals [14].

Bulls exhibit no symptoms and preputial cavity is mostly infected permanently as they grow older [15]. Hence using bulls below 4 years old is the simple control strategy [16]. Infected cows exhibit vaginitis. In pregnant animals it spreads to the cervix and uterus. Which leads to placentitis causing early abortion (1–16 weeks) or they exhibit unusual oestrous [11, 15]. Upon infection, healthy cows clear the infection within three months. The acquired immunity seemed to be short lived, usually for less than an year to the maximum of three years [15, 17].

### 2.1 *Diagnosics Methods for Detection of Trichomonas Foetus*

Disease can be suspected as a cause of reproductive failure based on symptoms like, irregularity in oestrous, early abortion, requirement of repeated mating. It was reported best to screen for the agent in preputial and vaginal washings or scrapings of the infected herd [16, 18]. Direct detection can be done by microscopy. Culturing using Modified Diamond’s medium further enhances the chances of direct detection [19, 20]. Rapid Giemsa-based staining techniques can also be used [18, 20]. For identification of *T. foetus* tissues like aborted foetus, fetal lungs and placenta were formalin fixed and Immunohistochemical methods using monoclonal antibodies were also used [21]. Nucleic acid detection methods are found to be sensitive. Hence PCR and RT PCR methods are developed using the highly conserved sequences of the “5.8 S ribosomal RNA gene and the flanking internal transcribed spacer regions” (ITS) as target regions [22–25].

A more suitable point of care diagnosis for nucleic acids is Isothermal Amplification Assay. A loop mediated, LAMP targeting *T. foetus* 5.8 S was developed and tested at the laboratory level [26]. This is a highly sensitive assay and requires minimal technical expertise. Such studies have to be further improved and tested at field level. As there is no systemic infection normally, immune responses to *T. foetus* was not reported in Bulls. Few tests were however developed like, mucus agglutination test and intra dermal test [21] and a most recently an antigen-capture enzyme-linked

immunosorbent assay. They could detect IgG antibodies in the mucinous vaginal discharge as well as in the serum of the infected cows. But less emphasis was laid on developing serology based diagnostics [15].

### 3 Limitations of the Available Diagnostic Assays

As per the OIE guidelines, Agent detection by culturing and microscopy is not specific, cumbersome, time taking, expensive and requires technical expertise. Molecular detection by PCR is expensive, requires sophisticated lab set up, technical expertise and the services are available only at few research centres like NDDDB, CDDL, RDDDL etc. compromising the application at resource limited stations at field level. Preputial washings and semen sample pose to get contaminated very easily interfere with the test results. Trichomonosis is very difficult to treat and hence Screening and isolations of infected animals is the only available strategy. Lack of diagnostic kits is leading to rapid spread.

#### 3.1 Significance of the Study

Transcriptomics and proteomics studies of *T. foetus* of help us to identify the set of highly expressed genes and proteins respectively. Screening such data and analyzing thoroughly would enable us find several putative antigenic proteins, comparative genomics based approach to identify putative antigenic proteins using proposed surface antigenic proteins of *T. vaginalis*, which is a widely studied human parasite causing infertility, would further enhance the process of identification of potential surface antigens of *T. foetus*.

Present proposal thus addresses to identify putative antigenic proteins of *T. foetus* which further pave way towards developing recombinant protein based diagnostic serological assays like ELISA and Lateral flow rapid kit for bulk screening of cattle in the field, organized farms, outbreaks, breeding stations as well as Individual in farms. Serological testing is simple, pose less contamination, rapid, inexpensive, sensitive, specific and easy to perform point of care diagnostics assays more suitable for field deployment. The potent surface antigenic proteins can also be further validated to develop Recombinant vaccine to control Trichomoniasis in cattle.

Till now, recombinant protein based diagnostics like ELISA and Lateral flow assays for screening Trichomonosis in cattle were not developed not only in INDIA but across the entire globe. Hence, it is a novel to screen transcriptomics and proteomics data and comparative genomics with the closely related species to identify potential antigenic proteins from *T. foetus* for enabling researchers to develop recombinant protein based point of care diagnostics.



## 4 Outline of the Approach to Identify the Potential Targets

Extensive literature survey in PubMed was carried out to search for Transcriptomics and proteomics studies of *T. foetus* and *T. vaginalis*. Screening and listing of highly expressed surface proteins, virulent proteins proposed to be involved in pathogenesis of Trichomonosis. Proposed potential virulent proteins of *T. vaginalis* are subjected to BLAST analysis to compare and identify the corresponding proteins of *T. foetus*. Obtained the accession numbers for corresponding highly expressed transcripts, ESTs and proteins in *T. foetus* and *T. vaginalis* from NCBI proteins data base. Downloaded the corresponding FASTA formats were downloaded.

Confirmation of listed proteins as potential putative surface proteins was done using TMHMM. FASTA sequences of the proteins were submitted to “TMHMM Server v. 2.0” for prediction of transmembrane helices in proteins. Output was obtained with graphics. Analyzed the FASTA sequences for the putative antigenic domains of the listed virulent proteins using “Epitope Prediction and Analysis Tools” in “Immune Epitope Database Analysis Resource” (IEDB). “B Cell Epitope Prediction Tools” were used to identify the putative antigenic domains of proteins, which are more likely to be recognized as epitopes with regard to antibodies produced by B cell response. The “BepiPred-2.0” server was used to predict B-cell epitopes from the FASTA sequence of the protein.

### 4.1 Important Finding and Analysis

In a study of “Functional profiling of the *T. foetus* transcriptome and proteome”, which is the first ever large-scale *T. foetus* expressed sequence tag (TfEST) project. When there was no complete genome sequencing data was available, sequencing ESTs gave the clue regarding the pathogenesis. Table 1 shows the list of ESTs coding for highly antigenic proteins [27].

Out of 15 Cysteine proteases expressed, 12 of these TfCPs were reported for the first time in this study. Comparison of amount of Cysteine protease of *T. foetus*, clearly demonstrates that Trichomonas produces large amount of CP8, followed by

**Table 1** List of highly expressed sequence tags in the cDNA library

S. No.	TfEST Contig No.	Highly antigenic transcripts
1	0742 12	Adhesin protein AP51-3
2	0752 15	Adhesin AP65-1like
3	0773 26	Adhesin AP65-1like
4	0781 36	Cysteine proteinase
5	0764 20	Cysteine protease
6	0776-2 23	Hypothetical protein1
7	0748 14	Hypothetical protein1

**Table 2** Putative immunogenic proteins *T. foetus* transcriptome analyzed by TMHMM for sub cellular localization

TfEST id	TfEST annotation
TfEST017A04	BspA-like leucine rich repeat surface antigen
TfEST006B04	Circumsporozoite protein precursor, putative
TfEST037G07	PPase: Ser/Thr Protein Phosphatase, PP2A catalytic subunit (PPN)
TfEST024F05	Similar to <i>T. vaginalis</i> hypothetical protein TVAG_235360
TfEST033A06	Calcium motive P-type ATPase, PMCA-type
TfEST073F08	Subtilisin-like serine peptidase
TfEST047E12	Ser/Thr Protein Phosphatase, Metallo-phosphoesterase
TfEST027H02	Protein SEY1, putative
TfEST_Contig0170	Tetraspanin
TfEST_Contig0461	Hypothetical protein TVAG_108330 (XP_001317322.1)
TfEST039E02	Hypothetical protein TVAG_247370 (XP_001580044.1)

CP16 and CP13. In the 2-Dimensional Electrophoresis analysis, spots corresponding to CP8 were abundant. Cysteine protease are found to be major antigenic proteins of *T. foetus* trophozoite (Table 2).

In an interesting study entitled “Comparative proteomic analysis of two pathogenic *T. foetus* genotypes: there is more to the proteome than meets the eye” the importance of CPs as predominant factors of pathogenicity and host-parasite interaction was showcased. It was also proposed to exploit CPs for diagnostic markers, vaccines as well as drugs development against *T. foetus* [28].

Another study titled “Comparative transcriptomics reveals striking similarities between the bovine and feline isolates of *T. foetus*: consequences for in silico drug-target identification” was conducted to understand the mechanism of adaption of Parasite host virulence in specific” [29]. CP8 is the most preferred transcript among all CPs. Apart from cysteine proteases, several other potential proteins were reported to be highly expressed in the bovine genotype. They have also reported several druggable domains, putative membrane proteins were also reported in Table 3.

**Table 3** Druggable domains of *Trichomonas foetus*

Transcript ID (c0_seq 1)	Transcript length	BlastX transcript identifier
Bc12_comp9993	3564	Pyruvate ferredoxin flavodoxin oxidoreductase
Bc12_comp9915	314	Clan family phytocystatin-like peptidase inhibitor
Bc12_comp5305	2198	gp63-like protein
Bc12_comp11217	921	Clan family metacaspase-like cysteine peptidase
Bc12_comp10022	973	Thioredoxin reductase

**Table 4** Unique pathogenic proteins identified in Costa fraction

NCBI entries	Entries proteins	TrichDB entries	Proteins	% Identities between <i>T. vaginalis</i> and <i>T. foetus</i> (%)
KX579561 KX579561	Adhesin AP65 1 precursor EC 1 1 1 40	KX579662	Malic enzyme	69
KX579602	CA family C19 ubiquitin hydrolase like cysteine peptidase EC 3 1 2 15	TVAG_475320	Clan CA, family C19, ubiquitin hydrolase-like cysteine peptidase	42
KX579645	Pyruvate: ferredoxin oxidoreductase A	TVAG_198110	Pyruvate-flavodoxin oxidoreductase,	66
KX579637	Clan CA family C1 cathepsin L like cysteine peptidase	TVAG_2980	Clan CA, family C1, Cathepsin L-like cysteine peptidase	64

The Costa protein rich undulating membrane present in *Trichomonas* was analyzed as the enriched fraction. Several unique proteins were identified to be involved in pathology are listed in the Table 4 [30].

*T. vaginalis* is closely related protozoan to *T. foetus* and exhibit similar pathology in human, more over since it is a human parasite, a lot of research work has been already done. Here we made an attempt to screen highly expressed antigenic proteins of *T. vaginalis* and with the idea of using comparative genomics based approach, identifying the related proteins in *T. foetus* [31]. In a review of *T. vaginalis* virulence factors, Bacteroides/spirochetes like surface protein A-like (BsPA) and Polymorphic membrane protein-like extracellular domain were found to be potential proteins [32].

In a study “Proteome analysis of the surface of *T. vaginalis* reveals novel proteins and strain-dependent differential expression” carried out to understand the mechanism behind tropism and survival of parasite in genital tract. six strains of *T. vaginalis* with differing adherence capacities were focused [33]. Several proteins like BspA-like, PMP-Like proteins and many proteases were highly expressed, with >2 fold increase in protein expression in more adherent strains. These proteins also exhibited high copy number in the genome.

Upon thorough analysis, few highly expressed most probable targetable proteins commonly found in *T. foetus* and *T. vaginalis* transcriptomics and proteomics were identified and shortlisted in this study in Table 5.

**Table 5** Brief summary of putative immunogenic proteins of *Trichomonas foetus*

Name of the protein	Accession No. <i>T. foetus</i>	No. of amino acids	Size of extracellular domain	Size of the largest antigenic domain
Cysteine protease 8	OHT13704.1	320	320	107
Surface antigen BspA-like Protein	OHS99292.1	3567	3498	476
Clan SB, family S8, subtilisin-like serine peptidase	OHT04136.1	866	819	69
Chlamydia polymorphic membrane protein-like extracellular domain	OHS93232.1	726	640	163
hypothetical protein TRFO_10391	OHS95735.1	870	837	216
Circumsporozoite protein precursor	OHT08051.1	241	241	25
Immuno-dominant variable surface antigen-like protein	OHT11175.1	270	270	30
Ser/Thr protein phosphatase	OHT00560.1	506	419	27
Tetraspanin family protein	OHS99351.1	208	192	113
Adhesin AP65	ARM19795.1	575	575	65
GP63-like protein	OHS97275.1	620	561	63

## 5 Conclusions and Future Prospects

Choosing putative immunogenic proteins indicated above, recombinant protein based diagnostic serological assays like ELISA and Lateral flow rapid kit can be developed for bulk screening of cattle in the field, organized farms, outbreaks, breeding stations as well as Individual in farms. The potent surface antigenic proteins can also be further validated to develop Recombinant vaccine to control Trichomonosis in cattle. Protein structure elucidation by X-ray crystallography and drug designing can be possible. Comparative genomics based projects are no cost, but has immense applications

## References

1. Compendium of Minimum standards of protocol & Standard operating procedures For Bovine breeding Government of India Ministry of Agriculture Department of Animal Husbandry, Dairying & Fisheries, India (2014)
2. Document of The World Bank, National Dairy Support Project. (2012)
3. Annual report department of animal husbandry, Dairying & Fisheries Ministry of Agriculture Government of India New Delhi (2012-13)
4. D.P. Rudrama, C.M. Prudhvi, P. Revathi, J. Mukta, N.M. Soumendra, S.S. Jagdip, P. Sudhakar, K.P.B. Kavi, P. Rathnagiri, Development and validation of rapid, sensitive and inexpensive protein g-based point of care diagnostic assay for serodiagnosis of paratuberculosis at resource-limited areas. *Current Trends in Biotechnology and Pharmacy*. **13**(3), 232–242 (2019)
5. R.C Mallikarjuna, V.K. Anumolu, C.M. Prudhvi, P. Revathi, M. Manasa, P. Rathnagiri, R. Vijayaraghavan, Sero prevalence and validation of in-house IgM elisa Kit for the detection of brucella antibody in human serum samples. *Asian J. Microbiol. Biotech. Env. Sc.* **19**(4), 975–980 (2017)
6. M. Manasa, P. Revathi, C.M. Prudhvi, V. Maroudam, P. Navaneetha, G.D. Raj, K.P.B. Kavi, P. Rathnagiri, Protein-G-based lateral flow assay for rapid serodiagnosis of brucellosis in domesticated animals. *J. Immunoassay Immunochem.* **40**(2), 232–242 (2019)
7. A. Warton, B.M. Honigberg, Structure of trichomonads as revealed by scanning electron microscopy. *J. Protozoot.*, **26**, 56–6 (1979)
8. R.M. Mariante, L.C. Lopes, M. Benchimol, *Tritrichomonas foetus* pseudocysts adhere to vaginal epithelial cells in a contact-dependent manner. *Parasitol. Res.* **92**, 303–312 (2003)
9. M.W. Gregory, B. Ellis, D.W. Redwood, Comparison of sampling methods for the detection of *Tritrichomonas foetus*. *Vet. Rec.* **127**, 16 (1990)
10. J.K. Elder, Examination of twelve strains of *Trichomonas foetus* (Reidmuller) isolated in Queensland and the description of a new serotype, *T. foetus* var. brisbane. *Queensl. J. Agric. Sci.* **21**, 193–203 (1964)
11. S.Z. Skirrow, R.H. Bondurant, Bovine trichomoniasis. *Vet. Bull.* **58**, 591–603 (1988)
12. J. Šlapeta, N. Müller, C.M. Stack, G. Walker, T.A. Lew, J. Tachezy, Comparative analysis of *trichomonas foetus* (Riedmüller, 1928) cattle genotype and *Tritrichomonas suis* (Davaine, 1875) at 10 DNA loci. *Int. J. Parasitol.* **42**, 1143–1149 (2012)
13. J. Tachezy, R. Tachezy, V. Hampl, M. Dinova, S. Vanacova, M. Vrlík, M. Van Ranst, J. Flegr, J. Kuldaa, Cattle pathogen *Tritrichomonas foetus* (Riedmuller, 1928) and pig commensal *Tritrichomonas suis* (Gruby & Delafond, 1843) belong to the same species. *J. Eukaryot. Microbiol.* **49**, 154–163 (2002)
14. C. Yao, Opportunistic human infections caused by *Tritrichomonas* species: a mini-review. *Clin. Microbiol. News* **34**, 127–131 (2012)
15. R.H. Bondurant, Pathogenesis, diagnosis and management of trichomoniasis in cattle. *Vet. Clin. North Am. Food Anim. Pract.* **13**, 345–361 (1997)
16. C. Yao, Diagnosis of *Tritrichomonas foetus*-infected bulls, an ultimate approach to eradicate bovine trichomoniasis in US cattle? *J. Med. Microbiol.* **62**, 1–9 (2013)
17. Trichomoniasis. OIE Reference manual—Chapter 3.04.15. Accessed on May 2019
18. N. Buller, B. Corney, Bovine Trichomonosis. Australian New Zealand Standard Diagnostic Procedure (2013). <http://www.agriculture.gov.au/SiteCollectionDocuments/animal/ah/ANZSDP-Bovine-trichomoniasis.pdf>. Assessed 1 Aug 2017
19. L.A. Bryan, J.R. Cambell, A.A. Gajadhar, Effects of temperature on the survival of *Tritrichomonas foetus* in transport, diamond's and InPouch™ TF media. *Vet. Rec.* **144**, 227–232 (1999)
20. Z.R. Lun, A.A. Gajadhar, A simple and rapid method for staining *Tritrichomonas foetus* and *Trichomonas vaginalis*. *J. Vet. Diagn. Invest.* **11**, 471–474 (1999)
21. J.C. Rhyan, K.L. Wilson, D.E. Bengess, L.L. Staokhouse, W.J. Quinn, The immunohistochemical detection of *Tritrichomonas foetus* in formalin-fixed paraffin-embedded sections of bovine placenta and fetal lung. *J. Vet. Diagn. Invest.* **7**, 98–101 (1995)

22. R.S.J. Felleisen, N. Lambelet, P. Bachmann, J. Nicolet, N. Muller, B. Gottstein, Detection of *Tritrichomonas foetus* by PCR and DNA enzyme immunoassay based on rRNA gene unit sequences. *J. Clin. Microbiol.* **36**, 513–519 (1998)
23. R.A. Grahn, R.H. Bondurant, K.A. Hoosear, R.L. Walker, L.A. Lyon, An improved molecular assay for *Tritrichomonas foetus*. *Vet. Parasitol.* **127**, 33–41 (2005)
24. L. Mcmillen, A.E. Lew, Improved detection of *Tritrichomonas foetus* in bovine diagnostic specimens using a novel probe-based real time PCR assay. *Vet. Parasitol.* **141**, 204–215 (2006)
25. D.C. Hayes, R.R. Anderson, R.L. Walker, Identification of trichomonadid protozoa from the bovine preputial cavity by polymerase chain reaction and restriction fragment length polymorphism typing. *J. Vet. Diagn. Invest.* **15**, 390–394 (2003)
26. J. Oyhenart, F. Martínez, R. Ramírez, M. Fort, J.D. Breccia, Loop mediated isothermal amplification of 5.8S rDNA for specific detection of *tritrichomonas foetus*. *Vet. Parasitol.* **193**, 59–65 (2013)
27. K.Y. Huang, J.W. Shin, P.J. Huang, F.M. Ku, W.C. Lin, R. Lin, W.M. Hsu, P. Tang, Functional profiling of the *tritrichomonas foetus* transcriptome and proteome. *Mol. Biochem Parasitol.* **187**(1), 60–71 (2013)
28. L.J. Stroud, J. Slapeta, M.P. Padula, D. Druery, G. Tsiotsioras, J.R. Coorssen, C.M. Stack, Comparative proteomic analysis of two pathogenic *Tritrichomonas foetus* genotypes: there is more to the proteome than meets the eye. *Int. J. Parasitol.* **47**(4), 203–213 (2017)
29. V. Morin-Adeline, R. Lomas, D. O’Meally, C. Stack, A. Conesa, J. Šlapeta, Comparative transcriptomics reveals striking similarities between the bovine and feline isolates of *Tritrichomonas foetus*: consequences for in silico drug-target identification. *BMC Genom.* **15**(1), 955 (2014)
30. A. de Ivone, I. Rosa, M.B. Caruso, E. de Oliveira Santos, L. Gonzaga, R.B. Zingali, A.T.R. de Vasconcelos, W. de Souza, M. Benchimol, The costa of trichomonads: A complex macromolecular cytoskeleton structure made of uncommon proteins. *Biol Cell.* **109**(6), 238–253 (2017)
31. K.Y. Huang, J.W. Shinc. P.J. Huangd, F.M. Kua, W.C., Lina, R. Lina, W.M. Hsua, P. Tanga, Functional profiling of the *tritrichomonas foetus* transcriptome and proteome. *Mol Biochem. Parasitol.* **187**, 60–71 (2013)
32. T.R.P. Hir, *Trichomonas vaginalis* virulence factors: an integrative overview. *Sex. Transm. Infect.* **89**(6), 439–443 (2013)
33. N. de Miguel, G. Lustig, O. Twu, A. Chattopadhyay, J.A. Wohlschlegel, P.J. Johnson, Proteome analysis of the surface of *Trichomonas vaginalis* reveals novel proteins and strain-dependent differential expression. *Mol. Cell. Proteomics MCP* **9**(7), 1554–1566 (2010)
34. National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov>
35. Immune Epitope Database <https://www.iedb.org/>
36. TMHMM Server v. 2.0 <http://www.cbs.dtu.dk/services/TMHMM/>

# SVM Based Approach to Text Description from Video Sceneries



Ramesh M. Kagalkar, Prasad Khot, Rudraneel Bhaumik, Sanket Potdar, and Danish Maruf

**Abstract** Human uses communication language either by written or spoken to describe visual world around them. The study of text description for any video goes increasing. This paper presents a system which produce English descriptions from the complex video samples. Here system produces text description from complex video, where it represents a framework that gives output as description for any long length video with multiple objects. This paper is broadly classified into two modules training and testing modules. Where the training module perform extracting of its unique features a with its description found in that video and is stored in database. In testing module consider the video sample which under goes frame extraction, preprocessing, segmentation, feature extraction and the extracted features are compared with features which are computed in training module then identify the video action, classify it and finally generate the text description using language model. The sentences are generated from objects for this assessment, a preferred database from youtube are accumulated in which 250 samples from 50 domain names. The performance of the system can be calculated and gives the accuracy of 90% with minimum processing time for object 2.

**Keywords** Text description · SVM classification · Video pre-processing and edge detection

---

R. M. Kagalkar · P. Khot (✉) · R. Bhaumik · S. Potdar · D. Maruf  
Department of CSE, KLE College of Engineering and Technology, Chikodi, Dist. Belagavi,  
Karnataka, India  
e-mail: [prasadkhot0@gmail.com](mailto:prasadkhot0@gmail.com)

R. M. Kagalkar  
e-mail: [rameshvtu10@gmail.com](mailto:rameshvtu10@gmail.com)

R. Bhaumik  
e-mail: [brudraneel@gmail.com](mailto:brudraneel@gmail.com)

S. Potdar  
e-mail: [sanketpotdar55@gmail.com](mailto:sanketpotdar55@gmail.com)

D. Maruf  
e-mail: [danishmaruf1234@gmail.com](mailto:danishmaruf1234@gmail.com)

## 1 Introduction

The computer vision has superior to identify the humans, categories their moves, or to differentiate between a large numbers of items and specify their attributes. The output is mostly a semantic representation encoding sports and gadgets categories. While such representations may be nicely processed by means of automatic systems, the natural way to talk this information with human beings is natural language. This system has a many applications within the area of human-computer interaction, producing precis descriptions of (internet) motion pictures, and automating film descriptions for deaf individuals.

The rest of this paper discuss is organized as follows. In Sect. 2, literature survey on video text description is presented. Section 3 depicts the system outline. The dataset used for experimental analysis in addition with predicted results were discussed in Sect. 4. And finally, conclusion has been discussed at the end.

## 2 Literature Survey

Inside the literature evaluation, a survey take a look at is completed on associated topics and some of the technique and troubles located are mentioned. In [1] paper deals with natural language access to video databases. Two techniques are proposed: within the first one queries are used to find the pictures similar to video key frames, and in the second one text content descriptions are generated from key frames and compare them with queries. In [2] this paper gives a new approach for text content segmentation in photographs of historical topographic maps and ground plans that uses a system mastering algorithm specialized in novelty detection to determine which additives of the picture are text. In [3] this paper, author cope with one of the most normal troubles of man or woman detection. The first one consists of a hierarchy of people, i.e., using the identify of various persons belonging to a collection for you to refine the person's detections. In [4] projected a scheme that gives a natural language description of images. The system consists a frame work of content planning and surface realization. To describe contents of image the technique they have used is vision detection and classification. In [5] gives a holistic data driven technique that generates a natural language description for videos. To describe the video content a subject verb object is used to generate the descriptive sentences. Similarly we refer the some of the limitation from different papers which are listed reference section [6–9].



### 3 System Outline

The system has mainly two modules, training and testing modules. In training module a set of input samples are consider and are stored in database. During training module, we consider input as a video samples for performing operation to extract its corresponding features and this features values are stored in database. Similarly perform for all the samples of database to extract its features and stored in database. In testing section considering an sample input sample and extract the features and then compare this features values with features which are already computed during training section. Also perform classification using SVM to recognize the text description and then generating text description. Figure 1 overview of system with block diagram.

#### 3.1 Training Module

In this module basically we consider a set of videos samples of different domains, then perform training to calculate features, then extracted features are stored in database. Similarly perform for all samples of database.

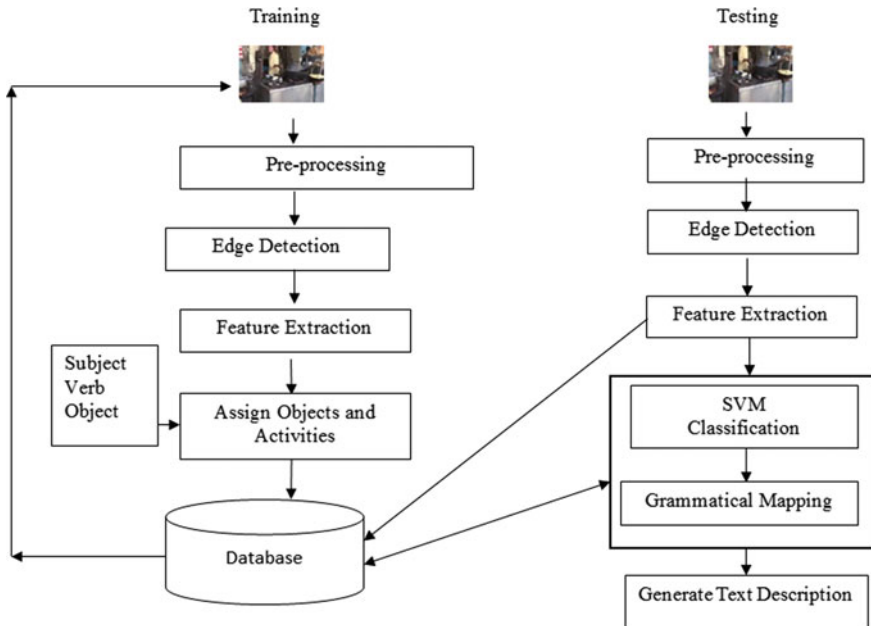


Fig. 1 Proposed system architecture

### 3.2 Testing Module

This module consider the testing sample as input and perform preprocessing, segmentation, extracting the features from the frames then it compare the extracted features values with database features values which are calculated during training module, then based on the features values we are going to identify and then classified to generate corresponding grammatically correct text description by comparing with database.

## 4 Database and Results Analysis

The dataset used for implementation require approximately 500 videos from various domain to evaluate text content extraction process, videos and its description are to be saved into the database to get first class effects. This videos are divided into numerous classes like shopping, street market place, water park, hospital, railway station, traffic signal, college area, airport etc. In this system we consider the set of samples for testing and is also listed in Table 1 for the consideration. Here consider each sample we need to calculate how much time it take, number of objets present in that video because it impact on the directly the sceneries of video. Then we have compare with

**Table 1** Results summary of presented system for objects 2

S. No.	Video samples domain	Time duration in seconds	Number of objects	Processing time in millisecond			Recognition rate in %
				Gaussian	Canny	SIFT	
1	Street market	31	2	1185	4430	25,100	Expected words = 17 Actual words = 15 Accuracy = 88%
2	Mall	17	2	593	8200	36,100	Expected words = 14 Actual words = 13 Accuracy = 92.8%
3	Playground	10	2	780	3900	3900	Expected words = 16 Actual words = 15 Accuracy = 93.7%

Gaussian, Canny and SIFT technique. Finally find the rate of recognition and also similarly carry out the remaining samples to compute.

The recognition rate is calculated based on the result generated. The video sample 1 expected word are 17 and the words generated by system are 15., So the recognition rate is 88%. Similarly, for mall and playground recognition rate calculated are 92.8 and 93.7%.

### 4.1 Analysis of Frame Extraction

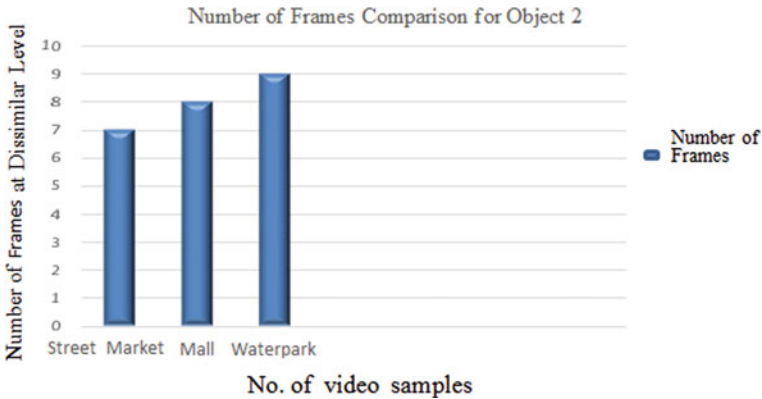
In Table 2 gives an number of frames comparison for object 2 are showed and Fig. 2 shows a graph for comparison of object 2.

Average number of frames extracted for object 2 =  $(7 + 8 + 9)/3 = 8$  frames.

So, the average number of frames extracted are 8 frames. Similarly, for object 3.

**Table 2** Number of frames comparison for object 2

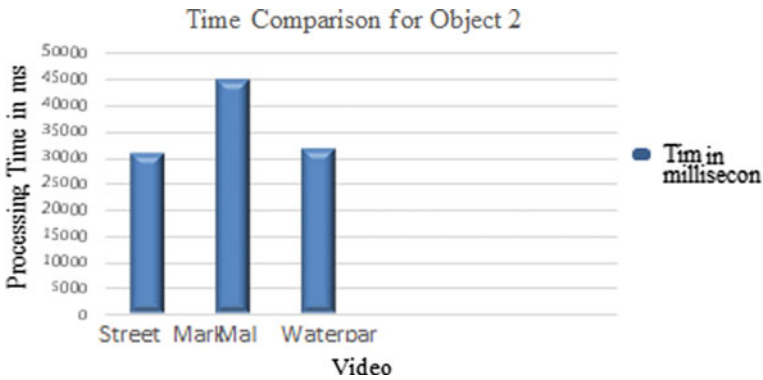
S. No.	Domain	Number of objects	Number of frames extracted
1	Street market	2	7
2	Mall	2	8
3	Water park	2	9



**Fig. 2** Graph for comparison of object 2

**Table 3** Time comparison for object 2

S. No.	Domain	Time in milliseconds
1	Street market	30,715
2	Mall	44,893
3	Water park	31,580



**Fig. 3** Graph for time comparison for object 2

### 4.2 Analysis of Processing Time

The total processing time taken by each video for object 2 is shown in Table 3 and the time for processing sample is depends on video size and number of object present in each frame. Figure 3 shows a graph for time comparison for object 2.

Average processing time for object 2 =  $(30,715 + 44,893 + 31,580)/3 = 35,729$  ms.

Therefore, the average time required for each video is 35,729 ms.

### 4.3 Analysis of Recognition Rate for Object 2

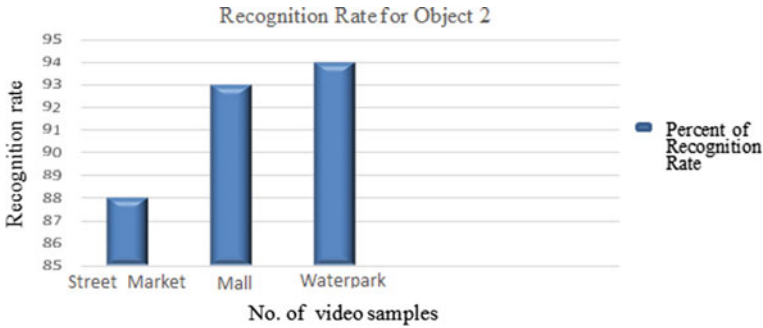
$$\text{Accuracy} = \frac{\text{Actual Words}}{\text{Expected Words}} * 100 = \frac{17}{15} * 100 = 88\%$$

So average recognition rate for object 2 =  $(88 + 93 + 94)/3 = 90\%$  (Table 4).

So the average recognition rate for video having object 2 is 90%. Similarly, for video sample mall and playground can be computed for future work. Figure 4 shows an graph of recognition rate for object 2. Here overall system performance can be measured in the number of processing time and recognition rate. The processing time depends on video size and number of object in each frame.

**Table 4** Recognition rate for object 2

S. No.	Domain	Recognition rate in %
1	Street market	88
2	Mall	92.8
3	Water park	93.7



**Fig. 4** Graph of recognition rate for object 2

## 5 Conclusion

The proposed system is an text description from video sample of two participants, it uses object detection, action recognition and features extraction from the given video sample and we perform preprocessing, segmentation and features extraction and extracted features are stored in database. Where in testing extracted features are compare with database features to classify, recognize to generate grammatically correct text description. The performance of the implemented system can be calculated and gives the accuracy of 93%. In future work, we will extend this paper work by considering large domain datasets with more number of participants of video samples for obtaining better accuracy and generate gramatical correct description.

## References

1. D. Francis, P. Pidou, B. Merialdo, B. Huet, Natural language access to video databases, in *IEEE Third International Conference on Multimedia Big Data* (2017)
2. S.C.S. Machado, C.A.B. Mello, Text segmentation in ancient topographic maps and floor plans with support vector data description, in *International Conference on Joint Neural Networks* (2015)
3. Á. García-Martín, R. Sánchez-Matilla, J.M. Martínez, Hierarchical detection of persons in groups. *Signal Image Video Process.* **11**, 1181–1188 (2017)
4. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, Siming Li, Y. Choi, A. C. Berg, T.L. Berg, BabyTalk: understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(12) (2013)

5. N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, S. Guadarrama, *Generating Natural-Language Video Descriptions Using Text-Mined Knowledge* (2013)
6. Barbu, A. Bridge, A. Burchill, Z. Coroian, D. Dickinson, S. Fidler, S. Michaux, A. Mussman, S. Narayanaswamy, S. Salvi, et al., Video in sentences out, in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 102–112 (2012)
7. C. Chang, C. Lin, LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 27 (2011)
8. Ding, D. Metze, F. Rawat, S. Schulam, P. Burger, S. Younessian, E. Bao, L. Christel, M. Hauptmann, Beyond audio and video retrieval: towards multimedia summarization, in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (2012)
9. M.U.G. Khan, Y. Gotoh, Describing video contents in natural language, in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data* (Association for Computational Linguistics, 2012), pp. 27–35

# Social Networking a Peril to Youth and Cultural Nuances—Needs Legal Fortification



G. Indirapriyadarsini and P. Neeraja

**Abstract** In the technological era, the world has become global village connecting people across the globe. Media is playing major role enabling the people to know many things that are happening in and around the globe instantly. Communication became easy with the advent of mobile phones and computers. However, there is vacuum of luxation and dissociation in practice. Hence the information communication technology replaced the human intelligence called as artificial intelligence.

**Keywords** Information · Technology · Legal provisions · Artificial intelligence · Human intelligence · Social networking

## 1 Introduction

Human being is a social animal hence it is obvious to maintain social relations. He or she cannot survive alone. It's evident since primitive societies that people live together in groups. In olden days' joint families were largely prevailing. They used to plan social gatherings in the name of festivals, birthdays, marriages, rituals, etc. to mingle with each other, get connected with people to share happiness and problems. These kinds of gatherings provide lot of scope for Children to play, share their feelings with the peer and have fun, relaxation, and excitement etc. Even during their stay at home elderly people use to teach them ethics and morals by way of stories and scientific reasons of our culture. All these practices help the children to imbibe ethics, sense of belongingness, responsibility, sharing, and logical thinking etc. There use to be prevalence of oneness. Each one feels the unity in diversity. There was a thought of all pervasiveness which leads to vasudaikakutumbam.

---

G. Indirapriyadarsini

Department of Law, Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India  
e-mail: [drindira36@gmail.com](mailto:drindira36@gmail.com)

P. Neeraja (✉)

Department of Women's Studies, Sri Padmavati Mahila Visva Vidyalayam, Tirupati, Andhra Pradesh, India  
e-mail: [neerajadws@gmail.com](mailto:neerajadws@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_54](https://doi.org/10.1007/978-3-030-46939-9_54)

With the people across the globe but still there is vacuum of luxation and dissociation in advent of technology in the era of globalisation, the world has become global village connecting practice. There is a development in transportation, communication and technology. Media started playing major role through print and electronic modes. Very instantly it enables the people to know many things that are happening in and around the globe with in no time. With the advent of mobile phones and computers, people could communicate with one another throughout the globe wherever they live. Hence the information communication technology replaced the human intelligence called as artificial intelligence.

### ***1.1 Evolution of Social Networking***

The first networking site developed in online was [www.classmates.com](http://www.classmates.com) in 1995. My space was found in 2006 consisting about 100 million user profiles. In 2003 for the first time Harvard started online facebook containing details of students. As man is fond of socialization and wants to be sociable social networking became more popular. People were more fascinated because it gives the ease in access for people to communicate. It is easy for people to write the required text and sends messages through texting, tweeting, and other social media's. Enthusiasts use blogs to express their ideas. This channel has become onset of entertainment. The outburst of equipment for taking photos, videos and development of technology to manipulate them lead to copyright violations and also became cause for child and women pornography. With the technological advances trade and commerce also has been initiated by online sites for example flipkart, amazon, ebay, etc. Increase in online trading resulted in E commerce contracts which again probe legal implications. The buyers are fascinated as they can sit at home and buy anything they require. The development of software's, programs, protections and many more has paved way for communication, entertainment and necessity. But all these were becoming possible for man could invent computers and other technology with the help of human intelligence. Researches who invented them also found that the skills exhibited by these computers are though narrow but some specialized skills were easier for computers like recognizing people, conversation with them, respond intelligently. Many artificial intelligence (AI) applications involve pattern recognition, speech recognition which was fanciful and sophisticated, now became common and necessity. The computer convinced man that it can replace the human subject and human intelligence. Hence we see the computers are used in the name of robots for various services instead of human beings. But this also brought to us problems, crimes are increased. Time pass entertainment is swallowing time and life especially of youth. The existing data relating to the number of internet using people in the year 2019 shows that India holds the second highest place and it was estimated to be 462.12 million, from these internet users about 258.27 million were likely to use social network [1].



## 2 Youth and Social Networking

Youth are backbone of nation. Healthy and skilful development youth lead to healthy and happy nation. But with the advent of technology and globalisation, the youth are fascinated by western culture losing human values, ignoring duties and misusing technology. Most of their times are used for browsing NET, watching social networking sites and also in playing video games. They chat with unknown faces in face book but ignore those who are in front of their faces. They forget the rest of the world and keep on watching social networking sites like LinkedIn, by tweeting in twitter, watching and sending messages in watts app, etc. The other kind of catch up from children to adults is video gaming. Relating to diagnosis of video gaming disorder world Health organization identified that after period of at least 12 months only it can be diagnosed, but early in cases of severity in signs and symptoms. It has recognized gaming disorder as major problem and has added as 11th one under International Classification of Diseases, i.e. disorders caused due to addictive behaviours with a code 6C51 [2]. Apart from radiation effects, visual and hearing problems, psychological problems, cognitive impairment, wasting their precious time, disabling their lives, video gaming is trapping youth and putting their life at risk.

## 3 Legal Provisions on Social Networking

Information Technology Act is the only legal protection available in India for regulating Social networking media. Social networking media are termed as “intermediaries” under the Indian Information Technology Act 2000 (IT Act 2000). The crux of IT Act relating to legal regulation of social networking media lies in Section 66A. In a landmark decision by Supreme Court in 2015 [3] struck down Section 66A of the Information & Technology Act, 2000 owing to unconstitutionality. It was held that the provision is violating Art. 19(1)(a) Freedom of speech provided under Indian Constitution. There is no legislation in India that regulates children’s access to social media or gives an express protection to them against cybercrimes. India’s cyber laws are largely governed by the IT Act, 2000. The IT Act Sections 66B and 67 states that, posting obscene content or pornography on social media is punishable. Posting pornographic material on social media may attract five years’ imprisonment and fine which may extend to Rs.10 lakhs [4].

The UNICEF India’s Child Online Protection in India Report, 2016 recognized the requirement for a legal framework to safeguard children from social media abuses. It stated, “Globally, online protection of children is most recognized and mentioned agenda however sadly India could be a very little late to understand it. It absolutely was discovered by UN agency that “Cyber-crimes against children square measure several forms like sex-texting, on-line grooming, production and distribution of harmful material for children, cyber bullying etc., are under-reported in India and have received little or no attention and aren’t enclosed within the National Crime Records

Bureau (NCRB) statistics as a separate category”. The UNICEF report additionally makes a robust case for separate legislation to guard children from social media abuse. Others who conducted surveys additionally urged for limiting or regulating children to access social media [5].

The central government has projected changes that aims at proscribing pretend news or rumors that reach to sizable amount with a one click on social media. It is abundantly essential to interrupt end-to-end encoding thus to determine place to begin point of the messages. “The social media platforms to deploy technology primarily based machine controlled tools or acceptable mechanisms, with appropriate controls, for proactively distinguishing or removing or disabling access to unlawful info or content”. As per the changes in law, the social media platforms can have to be compelled to befits the central government “within seventy-two hours” of a question. There ought to be a ‘Nodal person of contact for 24 × 7, coordination with enforcement agencies under law and officers to confirm compliance. The social media platforms are going to be keeping a vigil on “unlawful activity” for a time of “180 days” [6].

Further in its latest modification [7] to Information Technology Act new Section 67BA provides for penalty for publishing or pass on material repugnant to cultural attributes, Section 67BB regarding to Punishment for hosting dangerous on-line play resource, the section reads as Whoever hosts any online gaming resource which induces users to commit—(a) dangerous acts which are harmful to such users or others; or (b) acts which cause injury to themselves or others; or (c) any illegal act; he shall be punished on his first conviction itself with imprisonment for a term which may extend to one year and with fine which may extend to two lakh rupees. In the event of conviction subsequently for second time the penalty of imprisonment may extend to three years and additionally with fine which may extend to five lakh rupees. Chapter XIIB on Special Provisions regarding to on-line games is included in Section 79B. Whoever hosts an internet gaming resource or produces any storage media containing a gaming resource to be sold-out offline, shall ensure that—(a) the game resource is classified for use by suitable age group on the basis of game contents; and (b) there is a appropriate mechanism within the game resource to warn the users against repetition of the hazardous acts that are shown within the game in their real lives’ was inserted in 2018 [8].

However even after amendment, the identification and proof of game resources of hosts is troublesome, the game terms specify the age condition however it is often accessible for them, though children legally as per terms cannot play those games.

Public interest litigation has been filed before the Delhi High Court (HC) in which the amended WhatsApp’s privacy policy was challenged [8]. Judiciary during hearing Face book plea expressed its trepidation relating to the quick unfold of fake news, inducement of people on internet, assassination of the character. The Supreme Court stressed on the urgency for regulation of social media. The bench was headed by Justice Deepak Gupta made a significant observation on the mishandling of social media, worried about the working of some of the dangerous technologies. He also held that “With a little help he could have purchased an AK-47 on the Dark Web of on-line just in five minutes”. He recommended the government to vigilantly adopt

necessary measures and strict guidelines to regulate the social media; the court also upheld that the state while providing regulations must also consider on-line privacy, state sovereignty, the repute of an individual and recognize the need to balance both social media and social justice. The court also affirmed on the need to track originators of on-line crime and believed that if there is a technology to do such kind of illegal activity then there is a technology to prevent it also and nobody can escape with the excuse that there is no technology to decrypt social media content” [9]. “It could be a worldwide problem however there is a need to stumble on some solution. There ought to be some strict regulations and equally privacy of individuals should also be considered. This is the responsibility of policy makers. The bench also alleged that the social media is incredibly dangerous, the approach in which it is being misused” [10]. The Supreme Court observed that “the technology has taken a dangerous turn”. “TN government claimed that Facebook and other social media companies are not compliant with Indian laws, leading to amplified lawlessness and difficulties in ascertaining crimes” [11]. The press trust of India alleged that it is difficult for common man to understand the electronic communication app’s policy; terms of service which may also consists of deceptive terms [12].

Hence law ought to provide stringent measures and law enforcing agencies should strictly implement the provisions. The social media also needs to follow the rules strictly not violating legal provisions. Though there is legislative protection to some extent, judiciary has much concern about the problem, but the youth of the nation could not be brought out of ensnare. They could not prevent them to play dangerous games like pub G, blue whale, tik tok, etc.

## **4 A Study on Addiction of Social Network on Youth**

Globally people not merely use their leisure time for online networking but also spare extra time for social networking and video gaming which has become most trendy in current days. Excessive victimisation to mobiles, addicting to browse for networking sites and gaming are prone to cause increased physical and psychological health issues. This type of addiction is predominantly observed in students. Apart from health problems vital behavioural changes are noticed that lead to considerable blot in personal, family, social, educational, occupational relationships. In this perspective the researcher has undertaken the present study to identify the dominance of obsession to on-line video games and social networking in the youth [13].

### **4.1 Methodology**

A study was conducted by the researchers in September 2019 purposefully selected Tirupati as the area of research which being an educational hub. The universe of research was 120 young college going girls both undergraduate and post graduate

students age group of 19–22 years. The researcher purposively selected young adolescent girls for the study as youth are essential for development of nation. When youth are diverted from their actual achievements, their growth become diminutive and nation's development will be stunted. Empowering girls empowers nation.

Thence young girls are taken for the study. The researcher used Internet addiction test as it is found a reliable tool to measure obsessive use of internet. This tool is developed by Dr. Kimberly Young who is a psychologist and expert in internet addiction, which measures the presence and severity of Internet and technology addiction, which is nowadays viewed as a clinical disorder, requiring assessment and treatment. It consists of 20 items that measures mild, moderate and severe level of internet addiction [14]. The data was collected by researcher using Young's online gaming internet addiction scale questionnaire and collected data was analyzed by using SPSS 20.0 version.

## 4.2 Data Analysis

The researchers collected from college going girls and performed statistical analysis of the data collected and analysed the statistical data.

According to Internet addiction test total scores in each item are summed up, the level of addiction is measured as specified in the addiction test kit the higher the score, the greater level of addiction. If the score of respondents reads 20–49 points: they are identified as an average on-line user and surf the web a bit too long at times, but have control over the usage. If it scores 50–79 points: respondents are recognised as experiencing occasional or frequent problems because of the Internet and should consider the impact on their life. If it scores 80–100 points: The Internet usage is causing significant problems in the respondent's life and elevates the impact of the internet on their life and address the problems directly caused due to Internet usage' [15] (Table 1).

The analysis of data collected from young collage going girls has shown that the platforms which is most used for internet usage is through mobile phones, 60% of respondents are using mobile phones as their internet usage platforms. Only 40% of respondents are using other kinds of platforms for using internet. That too with the advent of Jio unlimited data provider, most of the respondents said that they are using jio unlimited data cards for internet access. The researcher studied social networking

**Table 1** Scoring of students

S. No.	Scoring points	Number of students	Percentage (%)
1	20–49	46	38.3
2	50–79	62	51.6
3	80–100	12	10
Total		120	100

**Table 2** Platform of internet usage of the respondents

Platform	Number	Percentage
Mobile phone	71	60.00
Laptops	34	28.00
Desktops	11	09.00
Internet centres	04	03.00
Total	120	100.00

and video gaming addiction using Young’s online gaming addiction scale. According to the present study 52% of respondents are having occasional disturbances, while, 10% students were experiencing severe disturbances in terms of sleep, concentration on studies, family relations and physical problems like headache, visual disturbances, etc. according to the scores that are scored by respondents (Table 2).

The researchers used Young’s online gaming addiction scale questionnaire to find out the respondents extent of addiction towards the social networking and video gaming using internet. As per the study it is evident that majority of them stay on-line longer than what they actually intended very frequently.

On interviewing and the scores that are notch up from the given questionnaire it was found that the average duration of using social networking sites and gaming was minimum of 3 h, while students who were addicted to them had used the devices maximum of 7 h. In the data analysis it was found that 38.3% of the students scored 20–49 points which shows that they are identified as an average on-line user and surf the web a bit too long at times, but have control over the usage. In the present study, the majority of respondent’s i.e., 51.6% scored 50–79 points which mean that the respondents are surfing online longer times and they are recognised as experiencing occasional or frequent health problems because of the Internet using mobile phones for not only browsing and checking emails but also for chatting in networking sites and playing games. It is observed that majority of the respondents are using internet frequently. Only 10% of girls scored 80–100 points that use internet extra longer times causing most significant health problems. Anyhow there was no statistical significance between undergraduate and postgraduate students regarding usage of internet and problems that are faced due to excessive use of these devices (Table 3).

## 5 Conclusion

The study has identified that students surfing internet to maximum extent. This increased obsessive use of internet lead to physical and psychosocial problems (problems like often bad moods, head ache, sleeplessness, anxiety, depression, be deficient in of social relations and poor in education). Hence there is urgent need in regulating the excessive use of Electronic devices to browse Internet for networking in social media and perilous video gaming. The youth are dragged to networking sites, attracted to video gaming which is putting their life at stake. Hence stringent rules

**Table 3** Information pertaining to respondent's extent of addiction towards internet [16]

S. No.	Question	Scale							Mean	SD
		1	2	3	4	5	0			
1	How often do you find that you stay on-line longer than you intended?	-	36	40	24	16	4			
2	How often do you neglect household chores to spend more time on-line?	8	16	44	36	8	8	20	16.4	
3	How often do you prefer the excitement of the Internet to intimacy with your partner?	8	12	60	40	-	-	20	16	
4	How often do you form new relationships with fellow on-line users?	4	4	92	20	-	-	20	24.53	
5	How often do others in your life complain to you about the amount of time you spend on-line?	4	28	60	20	8	-	20	36.04	
6	How often do your grades or school work suffers because of the amount of time you spend on-line?	16	28	32	44	8	8	20	22.2	
7	How often do you check your email before something else that you need to do?	4	32	40	28	8	8	20	15.75	
8	How often does your job performance or productivity suffer because of the Internet?	12	12	64	32	-	-	20	15.18	
9	How often do you become defensive or secretive when anyone asks you what you do on-line?	-	20	44	40	8	8	20	24.53	
10	How often do you block out disturbing thoughts about your life with soothing thoughts of the Internet?	14	21	40	32	4	9	20	18.24	
11	How often do you find yourself anticipating when you will go on-line again?	8	16	48	40	4	4	20	13.86	
12	How often do you fear that life without the Internet would be boring, empty, and joyless?	28	16	24	44	4	4	20	19.27	
13	How often do you snap, yell, or act annoyed if someone bothers you while you are on-line?	4	24	44	44	4	-	20	15.39	
14	How often do you lose sleep due to late-night log-ins?	24	12	32	44	4	4	20	20.4	
15	How often do you feel preoccupied with the Internet when off-line, or fantasize about being on-line?	-	4	48	64	-	4	20	16.2	

(continued)

**Table 3** (continued)

S. No.	Question	Scale					Mean	SD
		1	2	3	4	5		
16	How often do you find yourself saying “just a few more minutes” when online?	16	24	16	52	12	20	28.4
17	How often do you try to cut down the amount of time you spend on-line and fail?	20	40	20	32	4	20	17.53
18	How often do you try to hide how long you’ve been on-line?	14	28	44	25	4	20	12.47
19	How often do you choose to spend more time on-line over going out with others?	-	20	52	44	4	20	13.62
20	How often do you feel depressed, moody or nervous when you are off-line, which goes away once you are back on-line?	12	4	56	48	-	20	23.05

*Note* 0—does not apply, 1—rarely, 2—occasionally, 3—frequently, 4—often, 5—always

for placing any item in cyber space, before news placed or any game is placed in the cyber space it has to be checked for the security of children and youth and integrity of nation. Experts must use artificial intelligence (AI) in addition to human intelligence to identify the crime rather using it as an alternative source for human intelligence and using AI in the place of humans i.e., replacing humans. AI should advance and probe children and youth in developing and enhancing human intelligence rather addicting to devices and losing their cognitive development.

The children and youth in cyber era do not know the importance of culture and are carried away from what they watch in online and western cultural nuances, they forget their parents, people, place of birth and inbuilt human values. They work for money and money only without even have conscious that they are humans and suppose to imbibe certain values. The artificial intelligence should be tool in the hands of human beings for development but human intelligence for that matter man himself is in the grip of artificial intelligence violating basic rights of human beings, damaging the cultural nuances, addicting human brains and becoming peril to the development of youth. Hence in tune with technological development, there is need to promote more strict and suitable legal regulations such as censoring and licensing and stringent implementation measures and strategies to control online offences. The artificial intelligence shall be used more and more as a tool in identifying, detecting and presenting evidence in criminal justice system rather human beings becoming machines in the hands of artificial intelligence. AI should help as evidence in Justice System but cannot and should not substitute Judiciary. Especially children and youth who form future of nation should be protectively brought out of the influence of social networking and video gaming. Instead they should be made known the need to sharpen their brains, not depending on AI but on their own, importance of cultural nuances and social relations outside the cyber world and most important are human values.

## References

1. Sources: the Hindu
2. <https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1448597234>
3. Shreya Singhal and Ors. vs Union of India, 24 March, 2015, Bench: J. Chelameswar, Rohinton Fali Nariman (2015)
4. Information Technology Act 2000
5. P.K. Dutta, Safer Internet Day: Should India have law to regulate children's social media exposure?
6. India Today, New Delhi, 6 Feb 2018
7. The Information Technology [Intermediaries Guidelines (Amendment) Rules] (2018). [financialexpress.com](http://financialexpress.com)
8. Information Technology (amendment) Act (2018)
9. Gadgets Now News: WhatsApp privacy policy challenged in Delhi High Court: 5 things to know TNN | 15 Sept 2016. [timesofindia.indiatimes.com](http://timesofindia.indiatimes.com)
10. A. Chaturvedi, M. Mandavia: Give status of social media accountability norms: Supreme Court-Apex court will hear linking Aadhaar to Facebook a/cs, other tech platforms on Sept 24,



- says govt stand on accountability more important, ET Bureau, by Economic Times, Sep 14, 2019—visited on 20.10.19. Also by Indo-Asian News Service
11. A.A. Choudhary: SC wants rules for social media, gives gov. 24 Sept 2019. Read more at: [http://timesofindia.indiatimes.com/articleshow/71285287.cms?utm\\_source=contentofinterest&utm\\_medium=text&utm\\_campaign=cppst](http://timesofindia.indiatimes.com/articleshow/71285287.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst)
  12. Read more at: <https://yourstory.com/2019/09/supreme-court-social-media-misuse-facebook-whatsapp-twitter-india-government>
  13. By Press Trust of India
  14. Read more at: <https://yourstory.com/2019/09/facebook-indian-laws-tamil-nadu-aadhaar-supreme-court> (2019)
  15. P. Yarasani, et al.: Int. J. Community Med. Public Health **5** (10) (2018)
  16. Globaladdiction-Scales-InternetAddictionTest.pdf. <http://huibee.com/wordpress/wpcontent/uploads/2013/11/GLOBALADDICTIO-Scales-InternetAddictionTest.pdf>

# A Statistical Study on Analysing Repeated Measures of Data of Hyperlipidemia Cases



M. Siva Parvathi, K. Blessy Deborah, R. Vishnu Vardhan, T. Sukeerthi, and K. Sukanya

**Abstract** The present study explored the antihyperlipidemic potential of a standardized methanolic extract of *Averrhoa carambola* fruit in high fat diet-fed rats. *Averrhoa carambola* belonging to the family Oxalidaceae can be used as raw vegetable and ripe as fruit is used as traditional medicine. It possess high nutritional value and it is a good source of natural antioxidants. Although it is used in the treatment of diabetes, heart diseases and stroke due to presence of rich oxalates it is contraindicated in renal patients. Moreover, *A. carambola* was claimed and reported to have several ethanomedicinal uses like analgesic, anti-inflammatory, hypoglycaemic and hypocholesterolemic and so on. The presence of numerous phytoconstituents shows different pharmacological activities acting by different mechanisms. In addition, antioxidants are responsible for major treatments in curing ailments. The current study was aimed to evaluate the antihyperlipidemic activity of *A. carambola* in high fat diet induced hyperlipidaemia in male albino Wistar rats.

---

M. Siva Parvathi (✉)

Department of Applied Mathematics, Sri Padmavati Mahila Visvavidyalayam,  
Tirupati, Andhra Pradesh, India  
e-mail: [parvathimani2008@gmail.com](mailto:parvathimani2008@gmail.com)

K. Blessy Deborah

Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam,  
Tirupati, Andhra Pradesh, India  
e-mail: [blessypharma32@gmail.com](mailto:blessypharma32@gmail.com)

R. Vishnu Vardhan

Department of Statistics, Pondicherry University, Puducherry, India  
e-mail: [vrstatsguru@gmail.com](mailto:vrstatsguru@gmail.com)

T. Sukeerthi · K. Sukanya

Department of Statistics, Sri Padmavati Mahila Visvavidyalayam,  
Tirupati, Andhra Pradesh, India  
e-mail: [tptsukeerthi2309@gmail.com](mailto:tptsukeerthi2309@gmail.com)

K. Sukanya

e-mail: [sukanya.sudha27@gmail.com](mailto:sukanya.sudha27@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_55](https://doi.org/10.1007/978-3-030-46939-9_55)

**Keywords** *Averrhoa carambola* · Hyperlipidaemia · High-fat diet · Oxidative stress

## 1 Introduction

Hyperlipidaemia is a condition where excess of fatty substances called lipids, predominantly cholesterol and triglycerides are present in blood. It is also known as hyperlipoproteinemia due to the fatty substances in the blood attached to proteins. Coronary heart disease is a major cause of death in developed countries because of their sedentary lifestyle and various other factors. The role of lipids like total cholesterol, triglycerides, low density lipoprotein (LDL), VLDL, high density lipoprotein (HDL) in this disorder is extensively established. Plants were widely used medicines from the beginning of the early civilization in treating these kind of hyperlipoproteinemias.

*Averrhoa carambola* is small attractive tree belonging to the family of Oxalidaceae and it produces many fruits which are sour in taste but edible. The fruits of the plant were commonly known as star fruits in English due to its shape and having a wide range of ethno medical uses. The synthetic drugs available for the treatment of hyperlipidaemia are not showing good patient compliance of their cost and possible adverse effects.

Medicinal plants are used in treatment since early age of civilization. In fact westernized culture due to sedentary lifestyle 80% of the natural products are used for the ailments. *Averrhoa carambola* belonging to the family Oxalidaceae can be used as raw vegetable and ripe as fruit which has high nutritional value. From the time immemorial, the whole fruit is used as traditional medicine. *A. carambola* is widely grown in tropical and subtropical regions throughout the world and commonly called as star fruit due to its shape. It possess high nutritional value and it is a good source of natural antioxidants. Although it is used in the treatment of diabetes, heart diseases and stroke due to presence of rich oxalates it is contraindicated in renal patients.

Owing its rich nutritional value, antioxidant activity and claim of use in heart diseases, the present study is proposed to verify it, benefit as antihyperlipidemic agent.

## 2 Materials and Methods

### 2.1 Collection and Authentication of Plant Material

The fresh fruits of *A. carambola* were purchased from nearby fruit market in Tirupati, Chittoor Dist., Andhra Pradesh and authenticated by Prof. B. Sitaram, Sri Venkateswara Ayurvedic Medical College, Tirupati, Andhra Pradesh, India.

## **2.2 Preparation of *A. carambola* Powder**

The fresh fruits of *A. carambola* were purchased, washed and chopped into small pieces. Small pieces were shade dried and pulverised by means of a mixer to obtain fine powder.

10 g of *A. carambola* powder was weighed and 30 ml of distilled water was added to it. The mixture was boiled for 15 min to obtain aqueous extract. Preliminary phytochemical screening was performed for the aqueous extract of *A. carambola* powder as given by Kokate (2015) [1].

## **2.3 Chemicals**

*A. carambola* was purchased from the local market in Tirupati, Andhra Pradesh. The biochemical kits used were purchased from Erba Mannheim. All the other chemicals used in the present study were of analytical grade and purchased from Merck Ltd, India.

## **2.4 Experimental Animals**

Adult healthy male Wistar albino rats weighing 150–160 g were obtained from Sri Venkateshwara Enterprises, Bangalore. The animals were housed under standard environmental conditions like ambient temperature ( $25 \pm 10$  °C), relative humidity ( $55 \pm 5\%$ ) and a 12/12 h light dark cycle in polypropylene cages. Animals had free access to standard rodent-pellet diet and water ad libitum. All animal experiments were carried out in accordance with the guidelines of Committee for the purpose of Control and Supervision of Experiments on Animals. Institutional Animal Ethical Committee gave its approval to conduct the animal experiments and the approval No was 1677/PO/Re/S/2012/CPCSEA/IAEC/37 dated 23 Feb, 2019.

## **2.5 Dose Fixation**

The dose of *A. carambola* for the current study was fixed as 200 and 400 mg/kg, p.o as low and high doses respectively based on the earlier studies (Cazarolli et al. 2012) [2]. The high fat diet composition for inducing hyperlipidaemia was selected as per the reports of Srinivasan et al. (2005) [3].

## 2.6 Preparation of the Test Drug

Finely powdered *A. carambola* (ACP) was suspended in carboxy methyl cellulose (1% CMC W/V) everyday, freshly before administering to the animals. The test compound was administered by using oral feeding needle (18 G).

## 2.7 Composition of High Fat Diet

Vanaspati—40%, Vitamin and mineral mixture—20%, Casein—30%, Coconut oil—10%.

The above ingredients were mixed, 20% of the whole mixture was taken with 100 g of normal rat chow and dried. This diet was given to the rats everyday freshly throughout the treatment period in the form of pellets.

## 2.8 Experimental Design

The study animals were divided into five groups of five animals in each group. Rats were treated with high fat diet for 21 days to induce hyperlipidemia. The complete protocol was mentioned in the table.

Groups	Treatment	Purpose
I	Standard laboratory diet ad libitum	Serves as normal
II	High fat diet	Serves as disease control
III	High fat diet + standard (Atorvastatin 10 mg/kg, p.o)	Serves as standard
IV	High fat diet + low dose (200 mg/kg, p.o) of ACP	To assess antihyperlipidemic activity at low dose of ACP
V	High fat diet + high dose (400 mg/kg, p.o) of ACP	To assess antihyperlipidemic activity at high dose of ACP

## 2.9 Determination of Anthropometric Parameters

Body weights of the animals of all groups were measured every week. BMI's (waist to hip ratio) were calculated by dividing the individual animal weight with nose to anus length of the body in centimetres. Lee's index was calculated by dividing cube root of body weight with nose to anus length.

## 2.10 Biochemical Estimations

- (a) **Serum collection:** Blood was collected from the retro-orbital plexuses on day 0, 7, 14 and 21 from all groups. The collected blood samples were allowed to clot at room temperature and serum was separated by centrifugation at 2000 rpm for 15 min. Clear supernatant serum was collected to evaluate biochemical parameters like, Total cholesterol, Triglyceride, LDL-Cholesterol, VLDL-Cholesterol, HDL-Cholesterol were analyzed by using standard procedures as mentioned below.
- (b) **Collection of Liver samples:** At the end of the experimental study, animals were sacrificed by cervical decapitation. Then livers were carefully excised. The liver samples were fixed in 10% formalin solution for histopathological studies. The wet liver weight was noted to calculate the relative weight of liver.  $\text{Relative weight} = \text{Organ weight/body weight} \times 100$ .
- (c) **Histopathology:** Liver samples from various groups were fixed in 10% formalin and were processed for paraffin sectioning. Partial deparaffinization was done in hot air oven at 20 °C above the melting point of wax. Complete paraffinization of sections was done with hot xylene. The sections were dehydrated with ascending grade of alcohol and washed with distilled water. Then the sections were stained with hematoxylin and eosin and mounted with dibutyl phthalate xylene to study the histology of the liver sections. Evaluation was carried out under 10× magnification using Olympus microscope (X43).

## 3 Statistical Analysis

Data was expressed as mean  $\pm$  standard deviation (SD) of observations. Statistical difference was analysed repeated measures of using analysis of variance (ANOVA) followed by Tukey's multiple comparison test using SPSS Software Version 20 [4–9]. To depict mean difference a line chart is called Profile plot.

## 4 Results and Discussions

The data on AIP under each experimental group is collected on three different time points i.e., 7th Day, 14th Day and 21st Day. Such data is usually referred as repeated measures data. In this experiment, 5 groups are considered and three time points for procuring data, hence to analyse the data, the appropriate statistical tool is Repeated Measures ANOVA. Further, the pairwise comparisons across duration of the experiment and among 5 groups is carried out by Tukey's test. To depict the mean differences across the groups with respect to three points, a line chart by name *Profile plot* is prepared. All the results are compared at 0.05 level of significance.

The following table comprises of various statistical inputs such as descriptive summary of each group at each time and in total; the statistical significance values (p-values) of duration and groups and alphabets indicated as superscripts to notify the difference among 5 groups and duration. The superscripts depicted under total row, indicates that on average the AIP values at 7th day differ with 14th day but not at 21st day. It means that the response observed at 7th day and 21st day are almost similar and closer, whereas at 14th day, the mean AIP was observed to be lower than 7th and 21st days of experiment. Now, in terms of groups, a separate table is dedicated to explain the statistical significance. The pairwise comparisons reveals the insight that lower mean AIP is noticed in ‘Test High’ group and next is with ‘High Fat Diet’ and ‘Normal’ groups and further ‘Test low’ and highest mean AIP is with ‘Standard’ group. The mean AIP values of High Fat Diet and Normal are in same subset indicating they do not differ significantly, whereas, rest of the each differ significantly to each other. It is more evident with the profile plot.

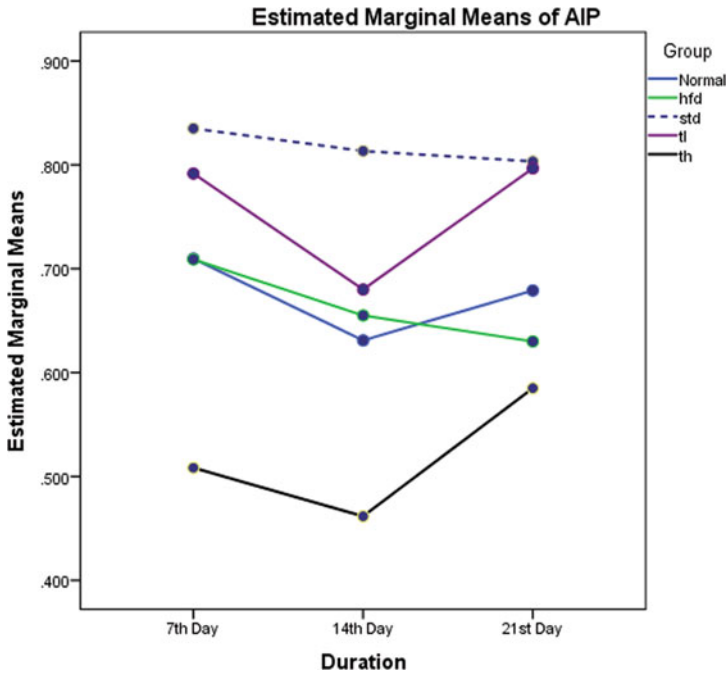
Group	AIP values at 7th day		AIP values at 14th day		AIP values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	0.7097	0.0654	0.6310	0.0504	0.6790	0.0425
High fat diet	0.7090	0.0470	0.6550	0.0266	0.6300	0.0358
Standard	0.8350	0.0302	0.8133	0.0242	0.8033	0.0266
Test low	0.7917	0.0299	0.6800	0.0610	0.7967	0.0602
Test high	0.5083	0.0366	0.4617	0.0483	0.5850	0.0647
Total	0.7107 <sup>a</sup>	0.1212	0.6482 <sup>b</sup>	0.1218	0.6988 <sup>a,c</sup>	0.0999

Duration: F-value = 1.454 (p-value: 0.239<sup>NS</sup>); group: F-value = 93.531 (p-value: 0.000\*)

Pairwise comparison using Tukey’s test				
Group	Subset			
	1	2	3	4
Test high	0.5183			
High fat diet		0.6647		
Normal		0.6732		
Test low			0.7561	
Standard				0.8172
Sig.	1.0000	0.9846	1.0000	1.0000

If we observe the pattern of each group in the profile plot, the ‘Test High’ groups’ mean AIP is lower at all 7th, 14th and 21st days and this is comparatively lower than the rest of the four groups, ‘High Fat Diet’, ‘Normal’, ‘Test Low’ and ‘Standard’

respectively. The mean AIP under ‘Standard’ groups are comparatively more than that of rest of the four groups.



In the table below, results show that there is no statistical significance across the different time points but on the whole, on an average, five variations of the groups observe to vary among them. This variations can be witnessed in the pairwise comparisons table, in which, similar mean BMI values are noticed with High fat diet and normal but they differ with the other three groups. Further, similar mean BMI is also noticed in Test low and Standard groups and these two attained higher average than that of the remaining groups. The same is depicted in the profile plot below.

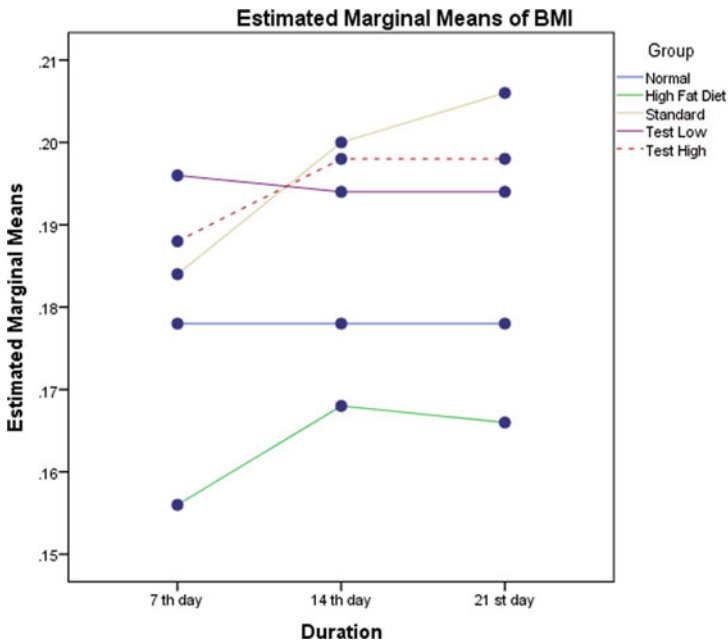
Group	BMI values at 7th day		BMI values at 14th day		BMI values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	0.1780	0.0084	0.1780	0.0084	0.1780	0.0084
High fat diet	0.1560	0.0207	0.1680	0.0217	0.1660	0.0241
Standard	0.1840	0.0167	0.2000	0.0316	0.2060	0.0371
Test low	0.1960	0.0182	0.1940	0.0114	0.1940	0.0055
Test high	0.1880	0.0045	0.1980	0.0130	0.1980	0.0084
Total	0.1804 <sup>a</sup>	0.0195	0.1876 <sup>b</sup>	0.0217	0.1884 <sup>b</sup>	0.0239

Duration: F-value = 3.030 (p-value: 0.097<sup>NS</sup>); group: F-value = 11.019 (p-value = 0.000\*)



Pairwise comparison using Tukey's test

Group	Subset		
	1	2	3
High fat diet	0.1633		
Normal	0.1780		
Test high		0.1947	
Test low			0.1947
Standard			0.1967
Sig.	0.163	0.090	0.997



In Body weight parameter, the significant results are noticed in both duration as well as across groups. The significance in duration means that there is gradual increase in the body weight from 7th day to 21st day (see the row with heading total). Further, in each and every group also, it is observed that on an average there is a hike. Using the Tukey's test, the pairwise combinations are taken and their statistical significance is noted and is reported in a separate table. Among the five groups, least average body weight is observed with 'Test Low' group and next to it are High fat Diet, Test High and Normal groups, whereas, noticeable growth in body weight is witnessed under the Standard group. From the profile plot, it is more evident that

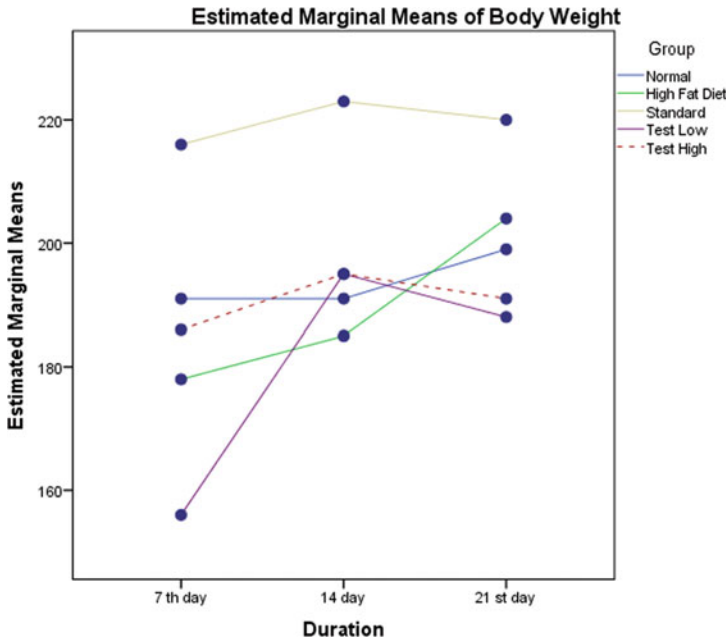
under Test Low group, the average body weight increased at 14th day and decreased in 21st day.

Group	Body weight values at 7th day		Body weight values at 14th day		Body weight values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	191.00	5.477	191.00	5.477	199.00	5.477
High fat diet	178.00	2.739	185.00	5.000	204.00	4.183
Standard	216.00	6.519	223.00	4.472	220.00	3.536
Test low	156.00	5.477	195.00	5.000	188.00	2.739
Test high	186.00	4.183	195.00	5.000	191.00	5.477
Total	185.40 <sup>a</sup>	20.357	197.80 <sup>b</sup>	14.148	200.40 <sup>b</sup>	12.241

Duration: F-value = 132.353 (p-value: 0.000\*); group: F-value = 111.143 (p-value = 0.000\*)

#### Pairwise comparison using Tukey's test

Group	Subset		
	1	2	3
Test low	179.67		
High fat diet		189.00	
Test high		190.67	
Normal		193.67	
Standard			219.67
Sig.	1.000	0.179	1.000

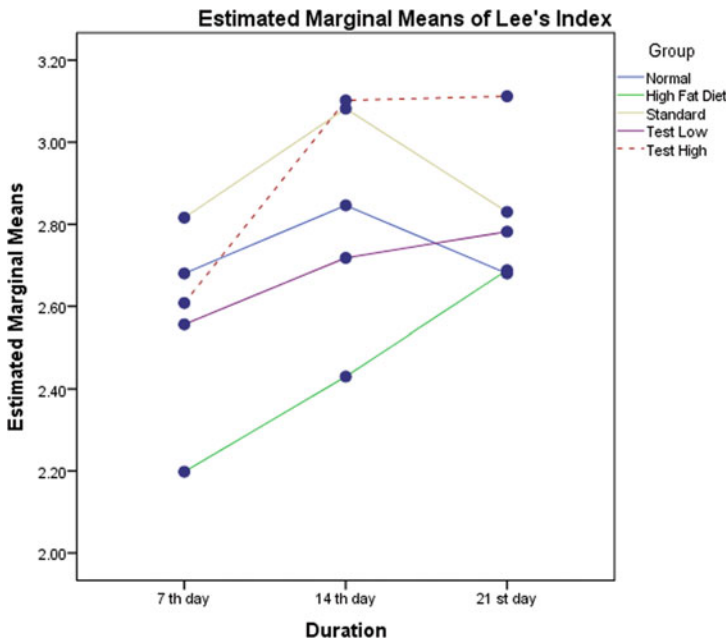


In the Lee’s Index parameter, the difference is noticed between 7th day and 21st day but not on 14th day. However, in this parameter also, statistical significance is noticed across groups. Similar Lee’s index is observed with High Fat Diet and Test Low samples whereas the other three groups have same Lee’s index. But the first two groups differ statistically with the later three groups.

Group	Lee’s index values at 7th day		Lee’s index values at 14th day		Lee’s index values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	2.6800	0.18960	2.8460	0.03647	2.6800	0.18960
High fat diet	2.1980	0.17050	2.4300	0.20603	2.6880	0.13646
Standard	2.8160	0.29203	3.0820	0.23584	2.8300	0.30553
Test low	2.5560	0.29712	2.7180	0.11862	2.7820	0.49231
Test high	2.6080	0.32175	3.1020	0.24004	3.1120	0.46575
Total	2.5716 <sup>a</sup>	0.31830	2.8356 <sup>a</sup>	0.30516	2.8184 <sup>b</sup>	0.35637

Duration: F-value = 1.454 (p-value: 0.239<sup>NS</sup>); group: F-value = 93.531 (p-value = 0.000\*)

Pairwise comparison using Tukey's test		
Group	Subset	
	1	2
High fat diet	2.4387	
Test low	2.6853	
Normal		2.7353
Standard		2.9093
Test high		2.9407
Sig.	00.110	0.092

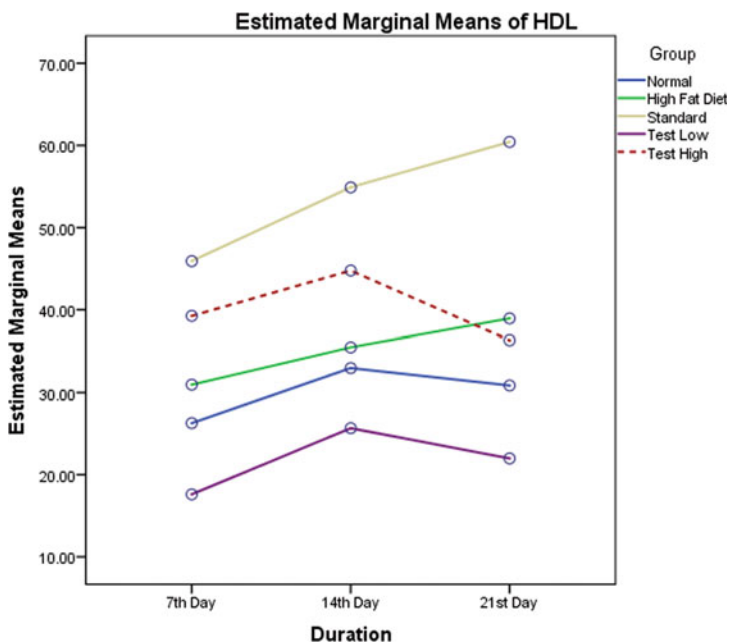


The mean HDL values were raised from 7th day to 14th day and remain static till 21st day. On comparing the groups on a pairwise basis, the least mean HDL was observed with Test Low samples, next gradual mean raise in HDL noticed with Normal, High Fat Diet, Test High and Standard respectively. In total, mean HDL of standard group is observed to possess higher values than the other groups.

Group	HDL values at 7th day		HDL values at 14th day		HDL values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	26.2733	3.63940	32.9750	3.66421	30.8550	1.95727
High fat diet	30.9550	1.28824	35.4850	2.20679	38.9583	2.49660
Standard	45.9250	4.50306	54.9017	3.55177	60.4333	5.51201
Test low	17.6150	2.80435	25.6583	3.44823	21.9817	2.84815
Test high	39.2583	3.07291	44.7683	4.58479	36.3233	4.97405
Total	32.0053 <sup>a</sup>	10.48294	38.7577 <sup>b</sup>	10.81841	37.7103 <sup>b</sup>	13.45710

Duration: F-value = 77.659 (p-value: 0.000\*); group: F-value = 148.453 (p-value = 0.000\*)

Pairwise comparison using Tukey's test					
Group	Subset				
	1	2	3	4	5
Test low	21.7517				
Normal		30.0344			
High fat diet			35.1328		
Test high				40.1167	
Standard					53.7533
Sig.	1.000	1.000	1.000	1.000	1.000



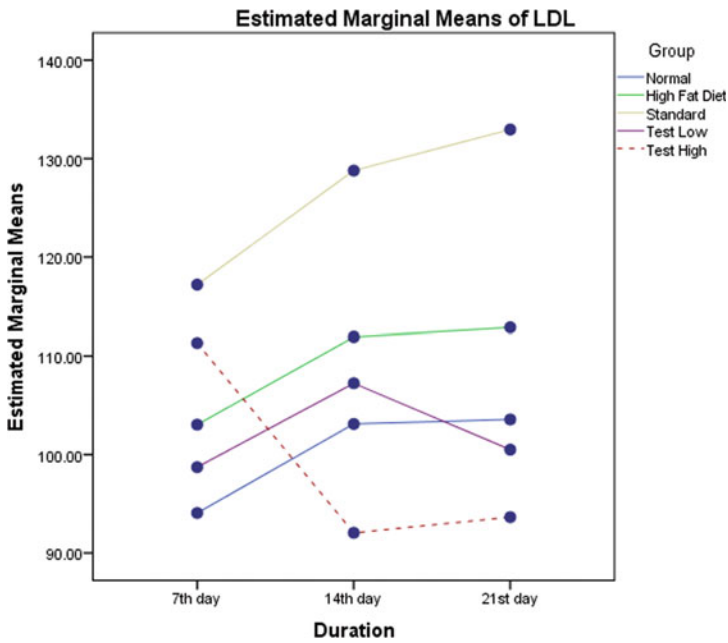
The mean LDL values were raised from 7th day to 14th day and remain static till 21st day. On comparing the groups on a pairwise basis, the least mean LDL was observed with Test High, Normal and Test Low samples, next gradual mean raise in LDL noticed with High Fat Diet and Standard respectively. In total, mean LDL of standard group is observed to possess higher values than the other groups.

Group	LDL values at 7th day		LDL values at 14th day		LDL values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	94.0583	3.87103	103.1100	5.07039	103.5700	4.60376
High fat diet	103.0367	5.96621	111.9233	5.72017	112.8683	4.36417
Standard	117.1933	5.63652	128.7817	3.22554	132.9550	3.63068
Test low	98.7217	2.80457	107.2417	3.64436	100.4983	2.32242
Test High	111.3183	6.14421	92.0467	4.48482	93.6500	3.54601
Total	104.8657 <sup>a</sup>	9.73590	108.6207 <sup>b</sup>	12.94176	108.7083 <sup>b</sup>	14.27748

Duration: F-value = 10.993 (p-value: 0.003\*); group: F-value = 93.381 (p-value = 0.000\*)

Pairwise comparison using Tukey's test

Group	Subset		
	1	2	3
Test high	99.0050		
Normal	100.2461		
Test low	102.1539		
High fat diet		109.2761	
Standard			126.3100
Sig.	0.341	1.000	1.000



The mean TC values were raised gradually from 7th day to 14th day and to 21st day. On comparing the groups on a pairwise basis, the least mean TC was observed with Normal samples, next gradual mean raise in TC noticed with Test High, Test Low, Standard and High Fat Diet respectively. In total, mean TC of High Fat Diet group is observed to possess higher values than the other groups.

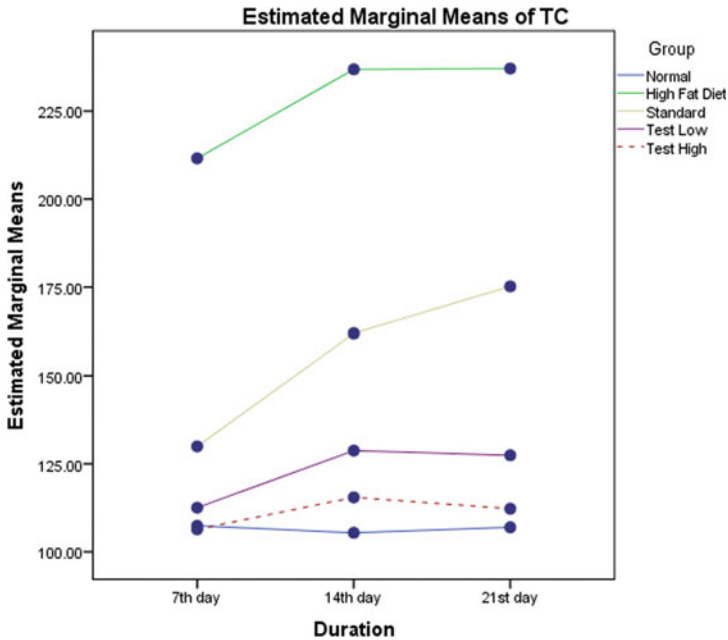
Group	TC values at 7th day		TC values at 14th day		TC values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	107.4083	1.97035	105.4117	3.13151	106.9600	3.28383
High fat diet	211.5683	2.64893	236.8383	3.58765	237.0750	3.66227
Standard	129.9367	2.61280	161.9600	6.06543	175.2150	5.34076
Test low	112.5450	5.14040	128.7467	4.51481	127.4167	5.66762
Test high	106.3933	3.98820	115.4883	4.97917	112.2500	5.11028
Total	133.5703 <sup>a</sup>	40.71442	149.6890 <sup>b</sup>	48.56890	151.7833 <sup>c</sup>	49.99709

Duration: F-value = 258.287 (p-value: 0.000\*); group: F-value = 3000.279 (p-value = 0.000\*)

Pairwise comparison using Tukey's test

Group	Subset				
	1	2	3	4	5
Normal	106.5933				
Test high		111.3772			
Test low			122.9028		
Standard				155.7039	
High fat diet					228.4939
Sig.	1.000	1.000	1.000	1.000	1.000





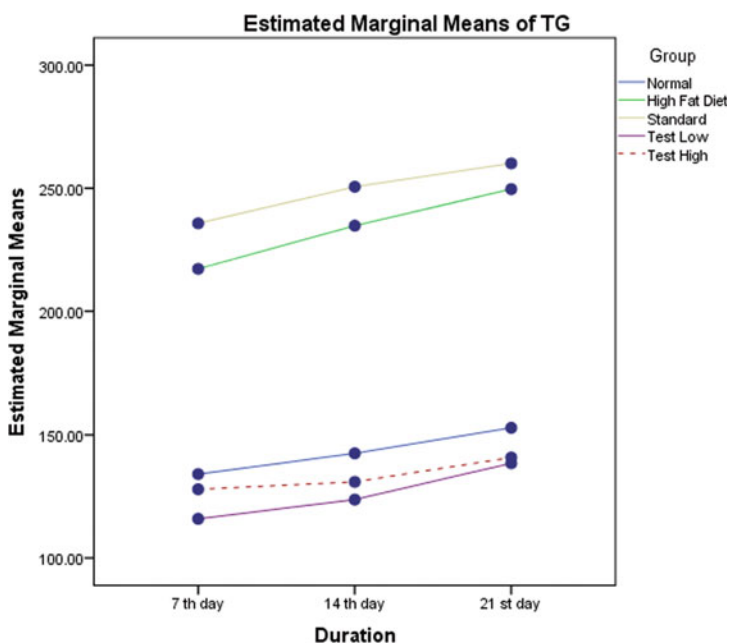
The mean TG values were gradually raised from 7th day to 14th day and to 21st day. On comparing the groups on a pairwise basis, the least mean TG was observed with Test Low samples, next gradual mean raise in TG noticed with Test High, Normal, High Fat Diet and Standard respectively. In total, mean TG of standard group is observed to possess higher values than the other groups.

Group	TG values at 7th day		TG values at 14th day		TG values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	134.0783	3.89881	142.5267	1.62370	152.9183	2.75405
High fat diet	217.2483	3.52164	234.7533	5.24523	249.6517	3.31379
Standard	235.7433	6.79515	250.5750	1.46725	260.1150	6.25462
Test low	115.9317	4.83326	123.7383	3.20930	138.4850	3.20030
Test high	127.9283	4.61467	130.9383	2.89982	140.8383	1.94683
Total	166.1860 <sup>a</sup>	50.98545	176.5063 <sup>b</sup>	55.59132	188.4017 <sup>b</sup>	55.64755

Duration: F-value = 413.564 (p-value: 0.000\*); group: F-value = 3550.493 (p-value = 0.000\*)

Pairwise comparison using Tukey's test

Group	Subset				
	1	2	3	4	5
Test low	126.0517				
Test high		133.2350			
Normal			143.1744		
High fat diet				233.8844	
Standard					248.8111
Sig.	1.000	1.000	1.000	1.000	1.000

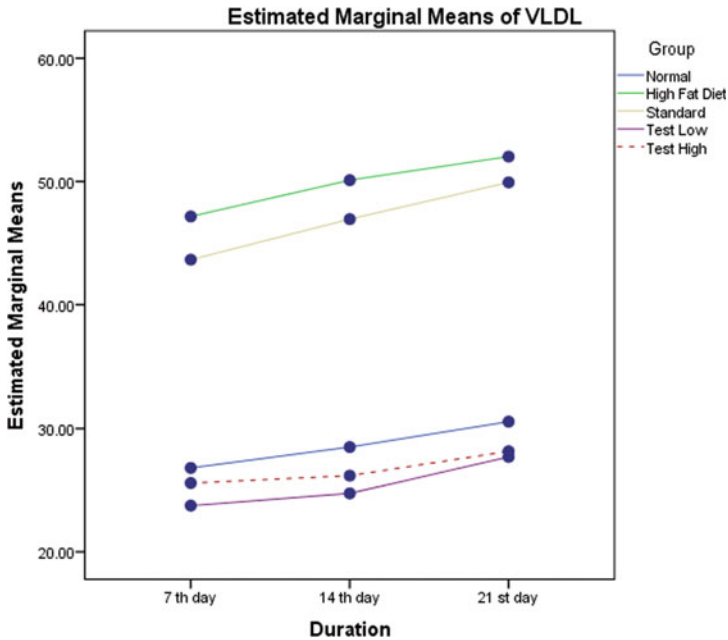


The mean VLDL values were gradually raised from 7th day to 14th day and to 21st day. On comparing the groups on a pairwise basis, the least mean VLDL was observed with Test Low samples, next gradual mean raise in VLDL noticed with Test High, Normal, High Fat Diet and Standard respectively. In total, mean VLDL of High Fat Diet is observed to possess higher values than the other groups.

Group	VLDL values at 7th day		VLDL values at 14th day		VLDL values at 21st day	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Normal	26.8117	0.7810	28.5017	0.3244	30.5683	0.5493
High fat diet	47.1600	1.3388	50.1100	0.2927	52.0183	1.2482
Standard	43.6550	0.8557	46.9433	1.0448	49.9267	0.6642
Test low	23.7483	1.7077	24.7433	0.6423	27.7033	0.6075
Test high	25.5817	0.9236	26.1833	0.5803	28.1650	0.3882
Total	33.3913 <sup>a</sup>	10.1500	35.2963 <sup>b</sup>	11.1173	37.6763 <sup>c</sup>	11.1284

Duration: F-value = 292.442 (p-value: 0.000\*); group: F-value = 2765.858 (p-value = 0.000\*)

Pairwise comparison using Tukey's test					
Group	Subset				
	1	2	3	4	5
Test low	25.3983				
Test high		26.6433			
Normal			28.6272		
Standard				46.8417	
High fat diet					49.7628
Sig.	1.000	1.000	1.000	1.000	1.000



## References

1. C.K. Kokate, A.P. Purohit, S.B. Gokhale, *Practical Pharmacognosy*, 2nd edn. (New Delhi, Vallabh Prakashan, 2009), pp. 466–470
2. L.H. Cazarolli, V.D. Kappel, D. F. Pereira, H. H. Moresco, I. M. C.Brighente, M. G. Pizzolatti, F. R. M. B. Silva, Anti-hyperglycemic action of apigenin-6-C- $\beta$ -fucopyranoside from *Averrhoa carambola*. *Fitoterapia*, 83(7), 1176–1183 (2012)
3. K. Srinivasan, B. Viswanad, L. Asrat, C. L. Kaul, P. Ramarao, Combination of high-fat diet-fed and low-dose streptozotocin-treated rat: a model for type 2 diabetes and pharmacological screening. *Pharmacol Res.* 52(4), 313–320 (2005)
4. L.S. Meyers, G.C. Gamest, A.J. Guarino, *Performing Data Analysis Using IBM SPSS*
5. G. Argyrous, *Statistics for Research: With a Guide to SPSS* (SAGE, London, 2005). ISBN 1-4129-1948-7
6. Alan Bryman, Duncan Cramer, *Quantitative Data Analysis with IBM SPSS 17, 18 and 19: A Guide for Social Scientists* (Routledge, New York, 2011). ISBN 978-0-415-57918-6
7. K.V.S. Sarma, *Statistics Made Simple, Do it Yourself on PC* (Prentice Hall of India, 2010)
8. J.J. Foster, *Data Analysis Using SPSS for Windows 8.0–10.0. Beginner’s Guide* (2001)
9. K. McCormick, J. Salcedo with Jon Peck and Andrew Wheeler, *SPSS statistics for data analysis and visualization*. May 2017

# Integrated Geospatial Technologies in e-Governance: An Indian Scenario



Pondari Satyanarayana, S. Jyothi, and Dandabathula Giribabu

**Abstract** Geospatial technology yields reliable location intelligence and in turn, location intelligence offers solutions for today's most pressing problems that are being faced by developing nations. Efficient public administration needs the knowledge of geographical information which ensures the proper delivery of government services. Integrated geospatial technologies comprise Earth Observation (remote sensing and meteorological observation) systems, GIS and spatial analytics (online and offline), Global or Regional Navigation Satellite System, and Geo-modelling services. Integrated geospatial technology can be plugged to e-Governance for efficient public services. This study presents the constituents of the integrated g-Governance toolkit, their application verticals, and the current trait of the Indian g-Governance system. Emphasis was given on the Indian geospatial platform namely Bhuvan and its contribution to effectiveness in performing the governance.

**Keywords** Electronic governance · Geospatial technology · Citizen centric services · webGIS

## 1 Introduction

Geospatial technology is a matured research area raised from the convergence and integration of various tools and techniques used to acquire and analyze spatial data in various research fields that include remote sensing, geographic information systems (GIS), geo-informatics, navigation systems, geography, geo-statistics, geophysics,

---

P. Satyanarayana (✉)  
National Remote Sensing Centre, ISRO, Hyderabad, India  
e-mail: [pondari@gmail.com](mailto:pondari@gmail.com)

S. Jyothi  
Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [jyothi.spmvv@gmail.com](mailto:jyothi.spmvv@gmail.com)

D. Giribabu  
Regional Remote Sensing Centre—West, NRSC, ISRO, Jodhpur, India  
e-mail: [dgb.isro@gmail.com](mailto:dgb.isro@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_56](https://doi.org/10.1007/978-3-030-46939-9_56)

and environmental science. Based on the development of integrated approaches and tools, a variety of location-specific information derived from multiple sources (optical sensors, microwave, lidar, GPS, citizen supplied, etc.) can be efficiently used for decision-making, problem-solving, emergency and management of disasters, and sustainable management of environmental resources [1].

The geospatial technology is the main strength of location intelligence which give answers to questions to pressing problems of current generation. The geographic locational knowledge facilitates effective people administration and guarantees adequacy in government service delivery. This has the advanced stage of geospatial technology, in United Nations geospatial technology is promoted to tackle the global challenges [2]. The incorporation of Geographical Information System (GIS), Earth Observation systems (EO), Global Position System (GPS), mobile mapping and cloud has recently increases the geospatial technology's usages capabilities.

The technologies and tools of geospatial has become an intrinsic part in our every-day life and become an indispensable and integral parts of day to day activities. Where 'awareness of location' is the main activity. Geospatial engineering, in example, the development of transportation industry is revolutionized for companies such as Lyft, Ola, Uber etc. [3]. In the similar way geospatial technology has played a crucial role in identifying global issues such as loss of the warming in globe and ozone layer. As far as government is concerned, location intelligence is a significant factor in functioning the every aspect and therefore GIS is a crucial component while intervention in the governments conceptualize, execute [4].

The country's uniform governance is not meet the citizen's needs and therefore problem solving approach with location intelligence as an key ingredient is need for hour. Information and communication technology (ICT) has penetrated nearly every area of modern life, including government services. Electronic governance or e-governance is the results from using technologies are improved for delivering and accessing government services for people's, employees and business partners benefits and it is created to reduce government operating costs, enhancing transparency and accountability.

In this paper explores the role of technology in geospatial is explored to enable public service performances in government. The constituents of geospatial like geo-services, geoportals and EO informations are formulated as a toolkit for government and its applications for serving the public after plugging into e-Governance. This article highlights current trends and characteristics of the discipline of g-Governance in India.

## **2 Constituents of Integrated Geospatial Technology**

The primary constituents of integrated geospatial technologies include EO systems, GIS infrastructure and data, data standards and data facilitators, global or regional navigation satellite system, Digital Elevation Models (DEMs) and etc. India's EO system has the capacity to serve the needs of changing government's. By EO data

and tools we can observe the appropriate positions, which will support wide range of utilization in the fields of water and land, atmosphere and ocean, eco-systems and environment, utilization in rural and urban and reduction in disaster risk [5]. National Remote Sensing Center (NRSC) is authorized for aerial and remote sensing satellite data acquisition, processing and dissemination in India (NRSC 2018). The national geo-portal is ISRO's Bhuvan ([bhuvan.nrsc.gov.in](http://bhuvan.nrsc.gov.in)) which is widely used by public, government, academia, non-governmental organizations [6].

Bhuvan has been established to address Indian satellite image dissemination, and theme-oriented services to allow stakeholder management and development activities to be scheduled, monitored and evaluated. Bhuvan platform provides ortho-corrected image base for entire Indian sub-continent, natural resource thematic datasets, Digital Surface Model (DSM), hydrological database (from basin to watershed) and millions of 'Points of Interest' data, geophysical products, host of customized government data and thematic services are collaboratively utilize tools and enabling g-Governance services.

In India, a satellite system is developed for regional navigation to get the exact location information to users in India as well as region extends up to 1500 km from its boundary. Which is an independent satellite named as Indian Regional Navigation Satellite System (IRNSS). The IRNSS space segment contains eight satellites in constellation, NAVIC, 3 satellites are located in suitable orbit slots in geostation orbit, 4 are located in geosynchronous orbit with equatorial crossing and inclination with two different planes. All the satellites are configured identically in constellation.

ISRO has launched 3 types of satellite series like and Oceanographic Atmosphere & Weather series (Oceansat & INSAT series), Resourcesat and RISAT series, Cartography (Cartosat series). These are generally referred for Water and Land Observation Systems of EO satellites (Table 1).

**Table 1** Constituents of Indian integrated geospatial technology

Entity	Description
Indian Regional Navigation Satellite System (IRNSS)	Regional navigation system dedicated to Indian sub-continent region using a combination of geosynchronous and geostationary satellites
Remote sensing satellites	ResourceSat-1, ResourceSat-2 OceanSat, and Cartosat series satellites providing imaging solutions to observe Land, Oceans and other ecosystems
Meteorological satellites	Meteorological satellites like INSAT-3D and INSAT-3DR contains imaging systems and atmospheric sounders

### 3 Applications of Integrated Geospatial Technology in Governance

g-Governance or Geospatial governance are referred as e-Governance geospatial plugin and acts as an extended e-Governance module with additional geospatial technology functionality [7].

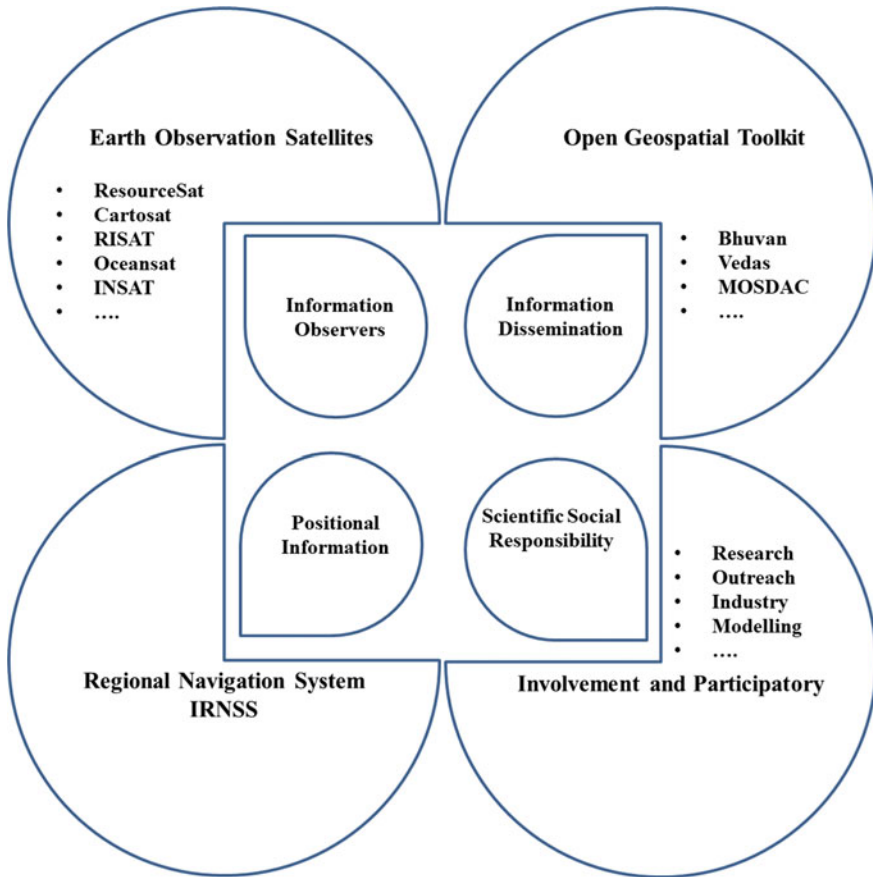
In recent times due to the advancements of EO systems, its applications have extended to improve human living conditions, growth of economy in social issues, and contribute the goals for development of sustainable [8]. The information of Spatial is derived from disaster risk reduction, communication, sustainable development, meteorological, navigation, EO synergy, ground-based observations plays a main part in the advantages of technology in space for effective management and socio-economic security. For a number of reasons, such as increasing inequality, globalization and the impacts of the financial crisis, as well as demographic transitions, social protection has become a focus of development policy. Geospatial technology help in providing enhanced access to social services in the most deprived and vulnerable regions, including those in remote areas with the help of synoptic view advantage. g-Governance helps in tracking and reviewing the progress of the planning and development activities, data disclosure and accountability issues. Openness, transparency, integrity, performance, and effectiveness are the key elements of good governance. Such elements will be ensured in good governance by an effective combination of ICT and geospatial resources along with an efficient strategy. g-Governance not only assists the government in obtaining dynamic environmental observations, but also helps to minimize the gap between science and society through citizenship. g-Governance results in ‘Landscape Governance,’ which views the landscape as a multi-functional, multi-stakeholder space and incorporates the climate, social and economic environment. Figure 1 shows few of application vertical or geospatial enabled governance activities.

Strategic applications in the areas of marine navigation, vehicle tracking, fleet management, Global Position System (GPS) integration in mobile phones, and etc. are possible standalone application of IRNSS. And as far as Indian EO systems are



Fig. 1 Verticals of government services with geospatial technology





**Fig. 2** Role of individual entities in the integrated geospatial technologies

concerned, it ensures the continuous available of data remote sensing data at various temporal, spectral and spatial resolutions on secure basis. Figure 2 shows the role of individual entities in the Integrated Geospatial Technologies.

#### **4 Applications of Integrated Geospatial Technology in Governance**

Federal ministries, state ministries, as well as line departments can take unfair advantages of the geospatial data and by using Bhuvan platform the services are consumed and accessed. Ministries concerned with Water Resources, Health & Family Welfare, Panchayati Raj, Sanitation, Education, Environment—Forest and Climate Change, Rural Development, Drinking Water Supply, Law & Justice, Communication and

Information Technology, Farmers Welfare and Agriculture, Home Affairs and Human Resources and Development State Forest Departments are some of the major users of Bhuvan platform.

Table 2 shows some of the initiatives taken by various ministries/department that consume the services of integrated geospatial technologies.

Ministry like Ministry of Communications, Ministry of Law and Justice, and etc. are using the Bhuvan platform for geotagging their infrastructure. Figure 3 shows the screenshot of Bhuvan web portal and Fig. 4 shows the geotagged assets for MGN-REGA project. Various ministries that deal with sustainable and inclusive growth of rural as well as urban India are effectively using the integrated geospatial technologies. Web services from the disseminators of integrated geospatial technology enable them to use the open data and provide good potential Governance.

**Table 2** Initiatives taken by various ministries/department to consume services of Indian Integrated Geospatial Technology

Ministry	Description
Farmers' Welfare and Agriculture Ministry	Chaman, Hailstorm, Horticulture, Pest Surveillance, Plantation, Pradhan Mantri Krishi Sinchayee Yojana and agro-meteorological services to project the Indian agriculture from climate change
Urban Development Ministry	Master plan formulation and smart cities
Ministry of Culture	Geo-legal advisory for heritage sites
Ministry of Water Resources	Geospatial information system for Namami Ganga project
Ministry of Panchayati Raj	SIS-DP and EPRIS
Ministry of Rural Development	GeoMGNREGA, IWMP, etc.
Forest and Environment Ministry	Environmental Information System (ENVIS), CRIS, Fly catchers
Human Resource & Development Ministry	NCERT curriculum related to geography
Ministry of Home Affairs	Census Information and Disaster Services
Urban Poverty Alleviation and Housing Ministry	Housing for all project
Ministry of Health and Family Welfare	National Health Resources Repository
Ministry of AYUSH	Information system for herbal gardens
State governments	Dedicated state government portals
Department of Forest (Karnataka, Himachal Pradesh, Uttarakhand)	Dedicated portals for forest information system
Water related portals	Walamtari, Water Body Information System, and etc.



Fig. 3 Bhuvan web portal screenshot for public

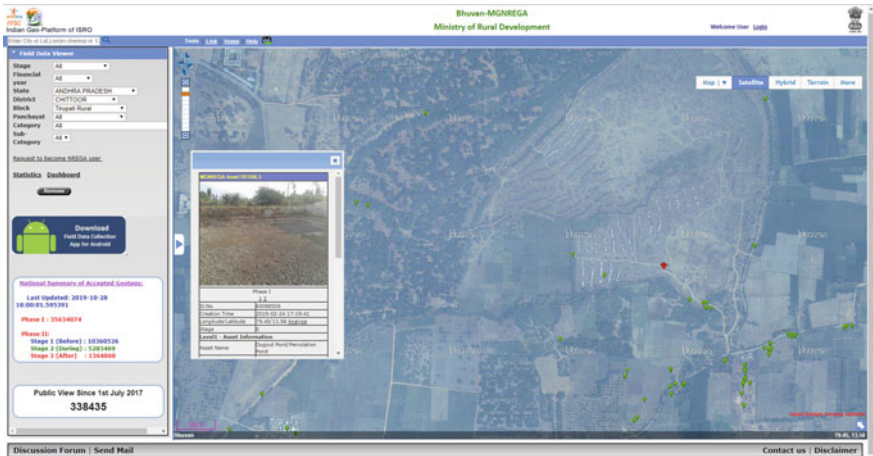


Fig. 4 Asset information of citizens are disclosed by using MGNREGA geoportal

## 5 Conclusion

Integrated geospatial technologies have enormous potential to improve the value chain in development projects and government interventions. The plug-ability of g-Governance in the e-Governance enabled major ministries to monitor, disclose and effective management of the interventions. Integrated geospatial technologies become a pivotal reference source that provides vital directions for effective governance with the help of regional navigation system, remote sensing, and GIS.

## References

1. P. Imperatore, A. Pepe (ed.), *Geospatial Technology: Environmental and Social Applications. BoD—Books on Demand* (2016)
2. UN-GGIM, United Nations Global Geospatial Information Management. <http://ggim.un.org/>
3. A.K. Laha, S. Putatunda, Real time location prediction with taxi-GPS data streams. *Transp. Res. Part C: Emerg. Technol.* **92**, 298–322 (2018)
4. R.C. Mathur, *From E-Governance To Geo-Governance—Geo Enabling For Better Governance* (egov, 2017)
5. R.K. Jaiswal, S. Bhatawdekar, Indian Earth Observation Program, in *Comprehensive Remote Sensing*, ed. by S. Liang (Elsevier, 2017)
6. Bhuvan, <https://bhuvan.nrsc.gov.in>
7. D. Giribabu, S.S. Rao, C.S. Reddy, P.V.P. Rao, Coordination with the help of geographical coordinates: g-governance in India. *J. Map Geogr. Libr.* **14**, 75–100 (2018)
8. M. Paganini, I. Petiteville, S. Ward, G. Dyke, M. Steventon, J. Harry, F. Kerblat, *Satellite Earth Observations in Support of the Sustainable Development Goals* (European Space Agency, 2018)

# Molecular Properties Prediction of N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-Substituted Phenylacrylohydrazides



K. Saritha and G. Rajitha

**Abstract** The substituted cinnamides were found as potent cyclooxygenase inhibitors, anti-convulsant, antioxidants, anti-inflammatory agents, analgesic, anti-microbial, anti-tubercular, antiviral, schistosomiasis, anti-platelet and anti-tumoral activities. We found it interesting to design substituted cinnamides, a series of N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-substituted phenylacrylohydrazides. Now-a-days QSAR studies has become an principal tool for designing more potent drugs. This plan uses Structural activity relationship, motif study of QSAR novel drugs to design selective inhibitor molecules using different softwares. The new motifs were predicted for their drug likeness and ADME studies. All the derivatives obeys Lipinski rule of five and has good bioactive scores.

**Keywords** Substituted cinnamides · Drug likeness · Bioavailability · In silico screening · Lipinski rule and bioactive scores

## 1 Introduction

### 1.1 Medicinal Importance of Substituted Cinnamides

Substituted cinnamides were described to possess variety of activities such as antioxidant [1], antimicrobial [2], antitumor [3], antitubercular [4], anti-inflammatory [5], antifungal [6], anticonvulsant [7] and are often used as promising precursor for the development of new, highly effective drugs. In silico tools offer many uses in giving chemical information and drug–target interactions. Web based tools used to

---

K. Saritha (✉)

Department of Pharmaceutical Chemistry, KVSRR Siddhartha College of Pharmaceutical Sciences, Vijayawada, Andhra Pradesh 520010, India  
e-mail: [sarithareddykalamalla@gmail.com](mailto:sarithareddykalamalla@gmail.com)

G. Rajitha

Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam (Women's University), Tirupati, Andhra Pradesh 517502, India  
e-mail: [rajitha.galla@gmail.com](mailto:rajitha.galla@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_57](https://doi.org/10.1007/978-3-030-46939-9_57)

641

calculate various molecular properties relevant to drug motifs which includes the prediction of ADME properties, BBB penetration and solubility. In view of this, it has been planned to design new substituted cinnamides bearing Piperonal moiety and screened for molecular prediction properties.

## 2 Methodology

Mol inspiration on line [8] tool was used to study molecular descriptors like TPSA Pka, number of hydrogen bond acceptors and donors. It also offers other software tools like cheminformatics which includes SMILES, SDfile converters, tautomer generators, fragmentation of molecules and QSAR properties. Mol inspiration tools are written in Java acts as a supportive platform for Molinspiration.

Lipinski rule of five [9] were checked by using Molinspiration. Drug-likeness scores were generated by using Mol soft website [10]. The drug-likeness scores were compared with the earlier reports [11] for their analysis.

## 3 Results and Discussion

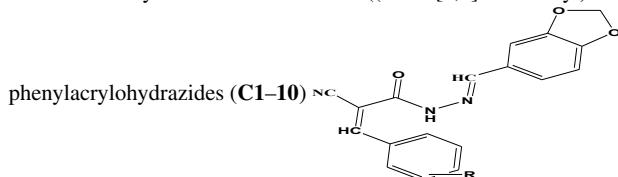
### 3.1 Design and Nomenclature

All the designed structural analogues shows good drug-likeness than the prototype with same pharmacophore essential for the activity. The fourteen compounds N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-substituted phenylacrylohydrazides (**C1–10**) were designed and their chemical structures were drawn using Marvin sketch. The electron donating like chloro, cyano and nitro and electron releasing groups like hydroxyl, methoxy, methyl etc. were tried on the benzylidene ring to study their effect on the predicted properties and bioactivities. Tables 1 and 2 show different substituted compounds and Nomenclature.

### 3.2 Prediction of Molecular Properties

Molecular properties for the designed compounds were predicted by using different softwares like Chemicalize, Mol soft and Mol inspiration. Lipinski's Rule of Five describes molecular properties important for a drug's pharmacokinetics [12].

All the synthesized compounds obeyed Lipinski's rule which includes hydrogen bond donors (less than 5), hydrogen bond acceptors (less than 10), rotatable bonds (less than 15), molecular weight (less than 500) and partition coefficient values (less than 5). The results obtained, given in detail in Table 3.

**Table 1** Physical data of N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-substituted

Compound code	R	Formula	O	Yield (%)
<b>C1</b>	C <sub>6</sub> H <sub>5</sub>	C <sub>18</sub> H <sub>13</sub> N <sub>3</sub> O <sub>3</sub>	123–125	70
<b>C2</b>	4-OH C <sub>6</sub> H <sub>4</sub>	C <sub>19</sub> H <sub>15</sub> N <sub>3</sub> O <sub>4</sub>	202–204	65
<b>C3</b>	4-OH,3-OCH <sub>3</sub> C <sub>6</sub> H <sub>3</sub>	C <sub>19</sub> H <sub>15</sub> N <sub>3</sub> O <sub>5</sub>	136–137	73
<b>C4</b>	4-OH, 3,5-(OCH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>2</sub>	C <sub>20</sub> H <sub>17</sub> N <sub>3</sub> O <sub>6</sub>	190–192	72
<b>C5</b>	4-OCH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	C <sub>19</sub> H <sub>15</sub> N <sub>3</sub> O <sub>4</sub>	140–141	74
<b>C6</b>	3,4-(OCH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>3</sub>	C <sub>20</sub> H <sub>17</sub> N <sub>3</sub> O <sub>5</sub>	191–192	68
<b>C7</b>	3,4,5-(OCH <sub>3</sub> ) <sub>3</sub> C <sub>6</sub> H <sub>2</sub>	C <sub>21</sub> H <sub>19</sub> N <sub>3</sub> O <sub>6</sub>	136–136	70
<b>C8</b>	4-N,N(CH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>4</sub>	C <sub>20</sub> H <sub>18</sub> N <sub>4</sub> O <sub>3</sub>	222–223	70
<b>C9</b>	4-Cl C <sub>6</sub> H <sub>4</sub>	C <sub>18</sub> H <sub>12</sub> ClN <sub>3</sub> O <sub>3</sub>	260–262	76
<b>C10</b>	4-CN C <sub>6</sub> H <sub>4</sub>	C <sub>19</sub> H <sub>12</sub> N <sub>4</sub> O <sub>3</sub>	215–216	78

Recrystallisation solvent: methanol

**Table 2** Nomenclature of the designed substituted cinnamides using Chemicalize.org

S. No.	R	Nomenclature
<b>C1</b>	C <sub>6</sub> H <sub>5</sub>	Smiles: N#CC(=Cc1ccccc1)C(=O)NN=Cc2cccc3OCOc23
<b>C2</b>	4-OH C <sub>6</sub> H <sub>4</sub>	Smiles: N#CC(=Cc1ccc(O)cc1)C(=O)NN=Cc2cccc3OCOc23
<b>C3</b>	4-OH,3-OCH <sub>3</sub> C <sub>6</sub> H <sub>3</sub>	Smiles: COc3cc(C=C(C#N)C(=O)NN=Cc1cccc2OCOc12)ccc3O
<b>C4</b>	4-OH, 3,5-(OCH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>2</sub>	Smiles: COc3cc(C=C(C#N)C(=O)NN=Cc1cccc2OCOc12)cc(OC)c3O
<b>C5</b>	4-OCH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	Smiles: COc3ccc(C=C(C#N)C(=O)NN=Cc1cccc2OCOc12)cc3
<b>C6</b>	3,4-(OCH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>3</sub>	Smiles: COc3ccc(C=C(C#N)C(=O)NN=Cc1cccc2OCOc12)cc3OC
<b>C7</b>	3,4,5-(OCH <sub>3</sub> ) <sub>3</sub> C <sub>6</sub> H <sub>2</sub>	Smiles: COc3cc(C=C(C#N)C(=O)NN=Cc1cccc2OCOc12)cc(OC)c3OC
<b>C8</b>	4-N,N(CH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>4</sub>	Smiles: CN(C)c3ccc(C=C(C#N)C(=O)NN=Cc1cccc2OCOc12)cc3
<b>C9</b>	4-Cl C <sub>6</sub> H <sub>4</sub>	Smiles: N#CC(=Cc1ccc(Cl)cc1)C(=O)NN=Cc2cccc3OCOc23
<b>C10</b>	4-CN C <sub>6</sub> H <sub>4</sub>	Smiles: N#CC(=Cc1ccc(C#N)cc1)C(=O)NN=Cc2cccc3OCOc23
11	Diclofenac	InChI key: DCOPUUMXTXDBNB-UHFFFAOYSA-N
12	Phenyl Butazone	InChI key: VYMDGNCVAMGZFE-UHFFFAOYSA-N

**Table 3** Prediction of molecular properties, bioavailability and drug-likeness score of N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-substituted phenylacryloylhydrazides (**C1-10**)

Compound	R	LogP <sup>a</sup>	LogS <sup>b</sup>	TSP A <sup>c</sup>	n-HBA <sup>d</sup>	n-HBD <sup>e</sup>	n-ROT B <sup>f</sup>	MW <sup>g</sup>	BA <sup>h</sup>	MUT <sup>i</sup>	TUM <sup>j</sup>	IRR <sup>k</sup>	REPE <sup>l</sup>	DL <sup>m</sup>
<b>C1</b>	C <sub>6</sub> H <sub>5</sub>	3.33	-5.07	83.72	6	1	4	319.32	0.55	G	G	G	G	Yes
<b>C2</b>	4-OH C <sub>6</sub> H <sub>4</sub>	2.85	-4.77	103.95	7	2	4	335.32	0.55	G	G	G	G	Yes
<b>C3</b>	4-OH,3-OCH <sub>3</sub> C <sub>6</sub> H <sub>3</sub>	2.67	-4.79	113.18	8	2	5	365.35	0.55	G	G	G	G	Yes
<b>C4</b>	4-OH, 3,5-(OCH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>2</sub>	2.68	-4.81	122.42	9	2	6	395.37	0.55	G	G	G	G	Yes
<b>C5</b>	4-OCH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	3.38	-5.09	92.96	7	1	5	349.35	0.55	G	G	G	G	Yes
<b>C6</b>	3,4-(OCH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>3</sub>	2.97	-5.11	102.19	8	1	6	379.37	0.55	G	G	G	G	Yes
<b>C7</b>	3,4,5-(OCH <sub>3</sub> ) <sub>3</sub> C <sub>6</sub> H <sub>2</sub>	2.96	-5.12	111.42	9	1	7	409.40	0.55	G	G	G	G	Yes
<b>C8</b>	4-N,N(CH <sub>3</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>4</sub>	3.43	-5.11	86.96	7	1	5	362.39	0.55	R	R	G	G	Yes

(continued)



Table 3 (continued)

Compound	R	LogP <sup>a</sup>	LogS <sup>b</sup>	TSP A <sup>c</sup>	n-HBA <sup>d</sup>	n-HBD <sup>e</sup>	n-ROT B <sup>f</sup>	MW <sup>g</sup>	BA <sup>h</sup>	MUT <sup>i</sup>	TUM <sup>j</sup>	IRR <sup>k</sup>	REPE <sup>l</sup>	DL <sup>m</sup>
<b>C9</b>	4-Cl C <sub>6</sub> H <sub>4</sub>	4.00	-5.81	83.72	6	1	4	353.76	0.55	G	G	G	G	Yes
<b>C10</b>	4-CN C <sub>6</sub> H <sub>4</sub>	3.08	-5.84	107.51	7	1	4	344.33	0.55	G	G	G	G	Yes

<sup>a</sup> Log P—partition coefficient

<sup>b</sup> Log S—solubility

<sup>c</sup> TPSA—topological polar surface area

<sup>d</sup> n-HBA—no. of hydrogen bond

<sup>e</sup> Acceptors

<sup>f</sup> n-HBD—no. of hydrogen bond donors

<sup>g</sup> n-ROTB—no. of rotatable bonds

<sup>h</sup> MW—molecular weight

<sup>i</sup> BA—bioavailability

<sup>j</sup> MUT—mutagenic

<sup>k</sup> TUM—tumorigenic

<sup>l</sup> IRR—irritant

<sup>m</sup> REPE—reproductive effect

<sup>n</sup> DL—drug-likeness

<sup>o</sup> G—no risk

<sup>p</sup> R—high risk

**Table 4** Bioactivity score of N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-substituted phenylacrylohydrazides (C1–10)

Compound code	GPCR ligand	Ion channel modulator	Kinase inhibitor	Nuclear receptor ligand	Protease inhibitor	Enzyme inhibitor
C1	-0.61	-0.83	-0.51	-0.66	-0.59	-0.43
C3	-0.54	-0.74	-0.44	-0.51	-0.54	-0.38
C4	-0.56	-0.76	-0.45	-0.57	-0.58	-0.37
C5	-0.54	-0.72	-0.42	-0.56	-0.54	-0.33
C6	-0.60	-0.82	-0.50	-0.64	-0.58	-0.43
C7	-0.57	-0.78	-0.47	-0.62	-0.57	-0.41
C8	-0.55	-0.74	-0.44	-0.63	-0.55	-0.39
C10	-0.55	-0.77	-0.43	-0.58	-0.56	-0.40
C11	-0.59	-0.80	-0.51	-0.66	-0.60	-0.45
C12	-0.54	-0.76	-0.39	-0.53	-0.53	-0.35

Topological surface area (TPSA) is also one of the key property that deals with bioavailability [13]. TPSA values for the tested compounds with less than 140 indicating their good oral bioavailability.

For all the synthesized compounds the bioactivity scores were calculated Table 4. If the bioactivity score is greater than 0 then the compounds are active, if  $-0.5$  to 0 then moderately active [14]. Anti-inflammatory and analgesic activity [15] are support by good kinase inhibitors and enzyme inhibition scores.

## 4 Conclusion

Rational drug design was based on computer based drug design for SAR studies. Literature survey explains the importance of substituted cinnamides and designed different N-((benzo[1,3]dioxol-5-yl)methylene)-2-cyano-3-substituted phenylacrylohydrazides and studied different properties which further indicate the probable potentiality of these compounds as future drugs.

## References

1. S. Gangadhara, C.H. Prasad, P. Venkateswarlu, Synthesis, antimicrobial and antioxidant activities of novel series of cinnamamide derivatives having morpholine moiety. *Med. Chem.* **4**, 778–783 (2014)
2. K. Saritha, G. Rajitha, K. SudheerKumar, A. Umamaheswari, Synthesis, molecular docking and antimicrobial activity of substituted cinnamides. *Int. J. Pharm. Biol. Sci.* **8**(3), 770–778 (2018)

3. D. Hadjipavlou-Litina, E. Pontiki, Aryl-acetic and cinnamic acids as lipoxygenase inhibitors with antioxidant, antiinflammatory and anticancer activity. *Methods Molecular Biology*, vol. 1208 (Humana Press, New York, NY, 2015), pp. 361–377
4. M.D. Kakwani, P.S. Vamsi, M.K. Ray, M.G. Rajan, S. Majee, A. Samad, M.S. Degani, Design, synthesis and antimycobacterial activity of cinnamide derivatives. *Bioorgan. Med. Chem. Lett.* **21**, 1997–1999 (2011)
5. P. Aikaterini, K. Dorothea, K. Christos, D. Hadjipavlou-Litina, Multitarget molecular hybrids of cinnamic acids. *Molecules* **19**, 20197–20226 (2014)
6. B. Korosec, M. Sova, S. Turk, N. Kravec, M. Novak, L. Lah, J. Stojan, B. Podobnik, S. Berne, N. Zupanec, M. Bunc, S. Gobec, R. Komel, Antifungal activity of cinnamic acid derivatives involves inhibition of benzoate 4 hydroxylase (CYP53). *J. Appl. Microbiol.* **116**, 955–966 (2014)
7. G. Li-Ping, W. Cheng-X, Q.D. Xian, S. Xin, P. Hu-Ri, Q. Zhe-S, Synthesis and anticonvulsant activity of N-(2-hydroxy ethyl) Cinnamide derivatives. *Eur. J. Med. Chem.* **44**, 3654–3657 (2009)
8. <http://www.molinspiration.com>
9. C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001)
10. <http://molsoft.com/mprop/>
11. G. Chen, S. Zheng, X. Luo et al., Focused combinatorial library design based on structural diversity, drug likeness and binding affinity score. *J. Comb. Chem.* **7**, 398–406 (2005)
12. D.F. Veber, S.R. Johnson, H.Y. Cheng, K.W. Ward et al., Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002)
13. P. Lalitha, S. Sivakamasundari, Calculation of molecular lipophilicity and drug likeness for few heterocycles. *Oriental J. Chem.* **26**, 135–141 (2010)
14. A. Verma, Lead finding from *Phyllanthus debelis* with hepatoprotective potentials. *Asian Pac. J. Trop. Biomed.* **2**, S1735–S1737 (2012)
15. H. Nirmala, A.R. Mullaicharam, Molecular modifications of ibuprofen using in silico modeling system. *Int. J. Nutr. Pharmacol. Neurol. Dis.* **2**, 156–162 (2012)

# Chitosan as a Heavy Metal Adsorbent in Waste Water Treatment



M. Saraswathi and R. J. Madhuri

**Abstract** Water pollution is the leading ecological complication in the earth particularly in India, which takes the millions of lives every year. 90% of leftover water pays the death of 2.2 million persons and 1.8 million youngsters per annum due to water-related diseases. For consumption pure and clean drinking water approachability is the primary right of human beings. Unfortunately, several toxic metals from different sources contaminate water sources which are leading to drinking water scarcity. Recently, chitosan, the derivative of chitin have been confirmed as a unique alternative adsorbent for toxic metals present in polluted water. Chitosan absorbs the great range of heavy metals depends on its crystallinity, attraction for water and degree of deacetylation.

**Keywords** Chitosan · Chitin · Heavy metals · Adsorption · Water pollution · Water treatment

## 1 Introduction

Water is the valued natural source, covering above 70% of the globe [1]. Without the apparently irreplaceable complex contained of hydrogen and oxygen, life on earth would be non-existent: it is necessary for everything on our planet to cultivate and flourish [2]. Out of 3% of freshwater available as surface water or groundwater only 0.06% is considered as potable [2]. Water is one of the more important resources for all activities of living things [3, 4].

Currently in several countries, approximately most of the population is unable to access pure drinking water for consumption [4]. Now-a-days water contamination is a severe problem by contaminating the water bodies very often by human activities, i.e., expansion in population, industrialization, urbanization, energy intensive life

---

M. Saraswathi · R. J. Madhuri (✉)  
Department of Applied Microbiology, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [drjayaravuri@gmail.com](mailto:drjayaravuri@gmail.com)

M. Saraswathi  
e-mail: [saraswathiphd@gmail.com](mailto:saraswathiphd@gmail.com)

© Springer Nature Switzerland AG 2020  
S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_58](https://doi.org/10.1007/978-3-030-46939-9_58)

style, deforestation, regulations and unprocessed sewage liberation [5–10]. And it also occurs when waste products are discharged into water bodies without acceptable treatment to get rid of detrimental compounds, resulting in contamination of drinking water [8, 11].

Huge concentrations of trace metal accumulation in natural water have several hazardous health effects on living things [12]. Trace metals interfere or inactivate the activity of the enzymes in living animals even in minute amounts. Because it is difficult to remove them completely from the environment once they enter in it [13]. About thirty metals play a vital role in trace amounts, for numerous biochemical and physiological mechanisms in living organisms as essential elements for existence [14]. Even those required as vital for several mechanisms, majority of the metals are venomous to living organisms if present in excess quantities [15].

In recent years, heavy metals are exhausting enormously to meet the augmented population requirements. Industries are discharging huge aggregates of nearly 20 heavy metals polluted waste water into environment. Heavy metals like Cadmium (Cd), Lead (Pb), Aluminum (Al), Silver (Ag), Nickel (Ni), Arsenic (As), Cobalt (Co) and Mercury (Hg) are frequently present in the contaminated water [16]. Because of their high solubility these heavy metals certainly enter into the aquatic ecosystems, food system of humans through consumption of aquatic animals [12, 17, 18].

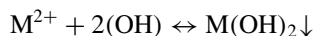
## **2 Different Methods Used to Eliminate Heavy Metals**

Adequate non-polluted freshwater is required for sustained existence of all living organisms and the maintenance of environmental balance. In view of the major human health impacts of noticeable intensities of heavy metals in water, throughout the world there have been continuous efforts from researchers for elimination of toxic compounds using different methods. Numerous procedures and technologies like chemical precipitation, ion exchange, solvent extraction, membrane processes, adsorption, coagulation and flocculation are the most frequently used techniques for heavy metal removal.

## **3 Different Methods Used to Remove Heavy Metal Contaminants**

### ***3.1 Chemical Precipitation***

It removes the heavy metals from contaminated water, in which added substances alter the physical state of the contaminants as insoluble precipitates [19–21]. Hydroxides (OH)<sup>-</sup> and sulfides (S)<sup>2-</sup> are commonly used chemical precipitants; precipitates heavy metals which are removed easily by filtration and sedimentation [10].



The drawback of this method is it needs huge quantity of substances to precipitate metals to an adequate level for discarded and the large capacity of precipitates involves additional treatment [10, 22].

### **3.2 Ion-Exchange**

Ion exchange is the utmost lavish technology, in which equally charged ions exchange occurs between trace metal ions and solid particle [17, 20]. Zeolites, or synthetically produced weak and strong anion and cat ion organic resins and chelating resins are usually used solid particles [23, 24]. The limitation of this method is it requires preprocessing of the contaminated water due to the huge accumulation of impurities.

### **3.3 Flotation**

It separates the solid phase and liquid phase by bubble attachment [19, 25]. The involved elements are disconnected from metal by the bubble enlargement [26]. Zeolite and chabazite are active accumulators with subtraction effectiveness of higher than 95% for an early metal concentration ranging from 60 to 500 mg/L.

### **3.4 Reverse Osmosis**

In RO a solution pass over a semipermeable membrane with compressed polymer matrix layer keeps the solute on one side and permits the clean solvent to pass to the other side [17, 21]. RO can take away numerous types of particles and ions. Synthetic water samples holding  $Cu^{2+}$  and  $Cd^{2+}$  ions at various concentrations are detached by RO with efficacy of 98% and 99% for  $Cu^{2+}$  and  $Cd^{2+}$ , respectively [27, 28].

### **3.5 Ultra Filtration**

Ultrafiltration membranes have pore size of nearly 0.002–0.1  $\mu m$  and MWCO around 10,000–100,000 Da and take away ions, microbiological species and dissolved solids from solution [30]. The marketable ultrafiltration membranes can receive feed water that carries high masses of impurities. About 18% of its total usages are for large industrial water treatment, and 15% of UF membrane is for waste water treatment.

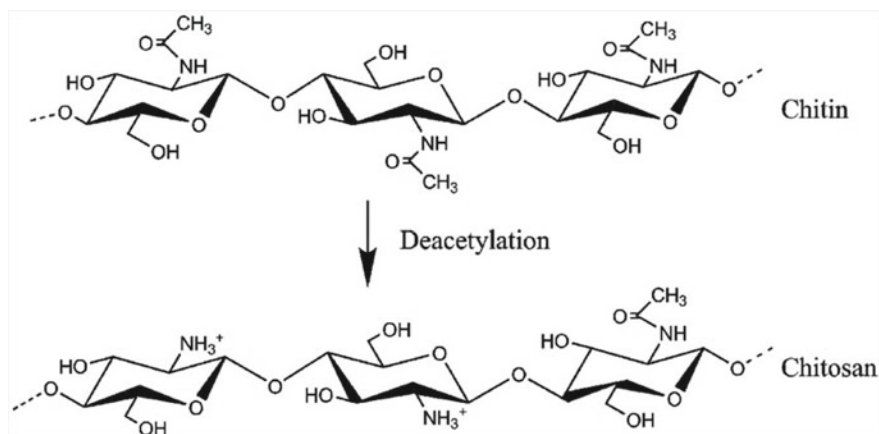
### 3.6 Adsorption

In adsorption the contaminants and poisonous metals present in the liquid phase are bound to the exterior of a solid by physical/chemical reactions. Among all above techniques this method seems to be more auspicious technology as it is low cost, needs low maintenance, economical and is energy proficient. Numerous adsorbents like silica gel, commercial zeolites, bauxite, xanthate, activated carbon, bentonite, lignin etc. are used to eliminate pollutants. In current years, natural resources confirmed as innovative resources removal of contaminants. Chitosan biopolymer an adsorbent is gaining more importance because of its availability, biodegradability and renewability [31].

### 3.7 Chitosan

Chitosan is a linear polysaccharide composed of randomly distributed  $\beta$ -(1  $\rightarrow$  4)-linked D-glucosamine and *N*-acetyl-D-glucosamine [32–34]. It is formed by alkaline *N*-deacetylation of chitin including deproteinization and deacetylation. Owing to many remarkable properties such as hydrophilicity, biocompatibility, biodegradability, non-toxicity, and existence of very reactive amino ( $-\text{NH}_2$ ) and hydroxyl ( $-\text{OH}$ ) groups in its backbone, chitosan has been used as an effective material for the elimination of toxic metals from untreated waters [35, 36] (Fig. 1).

The chitosan gained the multipurpose nature due to the presence of very sensitive groups like  $\text{C}_2\text{-NH}_2$ ,  $\text{C}_3\text{-OH}$ , and  $\text{C}_6\text{-OH}$ , respectively. Chitosan is present in crustaceans, mushrooms, fungi, etc. and produced mainly by alkaline deacetylation [37]. Chitosan and its byproducts have been extensively used in numerous fields



**Fig. 1** Structure of chitin and chitosan

including dispensary, biotechnology, cosmetics, nourishment, cultivation, ecological remediation; however, it has not yet been fully utilized on an industrial scale [38].

## References

1. L. Zhang, Y. Zeng, Z. Cheng, Removal of heavy metal ions using chitosan and modified chitosan: a review. *J. Mol. Liq.* **214**, 175–191 (2016)
2. S. Ahuja, Overview: sustaining water, the world's most crucial resource. *Chem. Water* 1–22 (2017)
3. J. Kostal, G. Giridhar Prabhukumar, U.L. Lao, A. Chen, M. Matsumoto, M. Ashok, W. Chen, Customizable biopolymers for heavy metal remediation. *J. Nanoparticle Res.* **7**, 517–523 (2015)
4. M.T. Amin, A.A. Alazba, U. Manzoor, A review of removal of pollutants from water/wastewater using different types of nanomaterials. *Adv. Mater. Sci. Eng.* **2014**, 1–24 (2014)
5. P.S. Verghese, M. Garg, Investigation of toxic heavy metals in drinking water of Agra city, India. *Orient. J. Chem.* **31**(3), 1835–1839 (2015)
6. A.M. Ahmed, M.R. Ali, A.K. Noor, Use of biocomposite adsorbents for the removal of methylene blue dye from aqueous solution. *Am. J. Mater. Sci.* **6**(5), 135–146 (2016)
7. S.B.A. Shawai, H.I. Muktar, A.G. Bataiya, I.I. Abdullahi, I.M. Shamsuddin, A.S. Yahaya, M. Suleiman, A review on heavy metals contamination in water and soil: effects, sources and phytoremediation techniques. *Int. J. Miner. Process. Extr. Metall.* **2**(2), 21–27 (2017)
8. M. Byambaa, E. Dolgor, K. Shiomori, Y. Suzuki, Removal and recovery of heavy metals from industrial wastewater by precipitation and foam separation using lime and casein. *J. Environ. Sci. Technol.* **11**(1), 1–9 (2018)
9. T.N. Batugedara, C.S.K. Rajapakse, Chitosan beads as a natural adsorbent for the removal of Cd (II) from aqueous solutions. *Int. J. Sci. Environ. Technol.* **6**(1), 606–619 (2017)
10. A. Azimi, A. Ahmad Azari, M. Rezakazemi, M. Meisam Ansarpour, Removal of heavy metals from industrial wastewaters: a review. *Chem Bio Eng Rev.* **4**(1), 1–2 (2017)
11. M.S. Islam, M.K. Ahmed, M. Raknuzzaman, M. Habibullah-Al-Mamun, M.K. Islam, Heavy metal pollution in surface water and sediment: a preliminary assessment of an urban river in a developing country. *Ecol. Indic.* **48**, 282–291 (2015)
12. M.A. Barakat, New trends in removing heavy metals from industrial wastewater. *Arab. J. Chem.* **4**, 361–377 (2011)
13. M. Mahurpawar, Effects of heavy metals on human health. *Int. J. Res.* 1–7 (2015)
14. M. Jaishankar, T. Tseten, A. Naresh, B.M. Blessy, N.B. Krishnamurthy, Toxicity, mechanism and health effects of some heavy metals. *Int. Discip. Toxicol.* **7**(2), 60–72 (2014)
15. M. Varsha, N. Madaan, A. Mudgal, R.B. Singh, S. Mishra, Effect of toxic metals on human health. *Open Nutraceuticals J.* **3**, 94–99 (2010)
16. L. Jarup, Hazards of heavy metal contamination. *Br. Med. Bull.* **68**, 167–182 (2003)
17. S.K. Gunatilake, Methods of removing heavy metals from industrial wastewater. *J. Multidiscip. Eng. Sci. Stud.* **1**(1), 12–18 (2015)
18. M.T. Alvarez, C. Crespo, B. Mattiasson, Precipitation of Zn (II), Cu (II) and Pb (II) at bench-scale using biogenic hydrogen sulfide from the utilization of volatile fatty acids. *Chemosphere* **66**(9), 1677–1683 (2007)
19. F. Fu, Q. Wang, Removal of heavy metal ions from wastewaters: a review. *J. Environ. Manag.* **92**(3), 407–418 (2011)
20. Renu, M. Agarwal, K. Singh, Methodologies for removal of heavy metal ions from wastewater. An overview. *Interdiscip. Environ. Rev.* **18**(2), 124–142 (2017)
21. P.P. Prabhu, B. Prabhu, A review on removal of heavy metal ions from waste water using natural/modified bentonite. *MATEC Web Conf.* **144** (2018)



22. U. Wingenfelder, B. Nowack, G. Furrer, R. Schulin, Adsorption of Pb and Cd by amine-modified zeolite. *Water Res.* **39**(14), 3287–3297 (2005)
23. N.H. Shaidan, U. Eldemerdash, S. Awad, Removal of Ni (II) ions from aqueous solutions using fixed-bed ion exchange column technique. *J. Taiwan Inst. Chem. Eng.* **43**(1), 40–45 (2012)
24. Y.N. Thakare, A.K. Jana, Performance of high density ion exchange resin (INDION225H) for removal of Cu (II) from waste water. *J. Environ. Chem. Eng.* **3**(2), 1393–1398 (2015)
25. X. Wang, Y. Du, H. Liu, Preparation, characterization and antimicrobial activity of chitosan–Zn complex. *Carbohydr. Polym.* **56**, 21 (2004)
26. P. Jokela, P. Keskkitalo, Plywood mill water system closure by dissolved air flotation treatment. *Water Sci. Technol.* **40**(11–12), 33–41 (1999)
27. A. Subramani, J.G. Jacangelo, Treatment technologies for reverse osmosis concentrate volume minimization: a review. *Purif. Technol.* **122**, 472–489 (2014)
28. Metcalf, Eddy, *Wastewater Engineering: Treatment and Reuse*, 4th edn. (McGraw-Hill, New York, 2003)
29. S.B. Thaçi, T.S. Gashi, Reverse osmosis removal of heavy metals from wastewater effluents using biowaste materials pretreatment. *Pol. J. Environ. Stud.* **28**(1), 337–341 (2019)
30. P. Azimi, P. Derakhshi, K. Tahvildari, F. Motiee, Application of chitosan absorbent in reducing the amount of airport wastes containing ethylene glycol pollution. *Euras. J. Biosci.* **13**(2), 781–789 (2019)
31. Rahmi, Marlina, Nisfayati, Comparison of cadmium adsorption onto chitosan and epichlorohydrin cross linked chitosan/eggshell composite. *Mater. Sci. Eng.* **352** (2018)
32. K. Bhavani, E. Roshan Ara Begum, S. Selvakumar, R. Shenbagarathai, Chitosan—a low cost adsorbent for electroplating waste water treatment. *J. Bioremediat. Biodegrad.* **7**(3), 1–6 (2016)
33. A. Agarwal, Vaishali, Chitosan based adsorbent: a remedy to handle industrial waste water. *Int. J. Eng. Sci.* **6**(9), 34–49 (2017)
34. A.B. Olohigbe, O.R. Etiosa, O. Okiei Wesley, Highly deacetylated chitosan as low-cost adsorbent material for removal of heavy metals from water. *Asian J. Phys. Chem. Sci.* **5**(2), 1–7 (2018)
35. J. Wang, S. Zhuang, Removal of various pollutants from water and wastewater by modified chitosan adsorbents. *Crit. Rev. Environ. Sci. Technol.* **47**, 2331–2386 (2018)
36. F.O.M.S. Abreu, N. Alves da Silva, M. de Sousa Sipaubá, T.F.M. Pires, T.A. Bomfim, O. Aze, Chitosan and gum arabic nanoparticles for heavy metal adsorption. *Polímeros* **28**(3), 231–238 (2018)
37. A. Rafique, K.M. Zia, M. Zuber, S. Tabasum, Chitosan functionalized poly (vinyl alcohol) for prospects biomedical and industrial applications: a review. *Int. J. Biol. Macromol.* **87** (2016)
38. M. Rinaudo, Chitin and chitosan: properties and applications. *Prog. Polym. Sci.* **31**(7), 603–632 (2006)
39. N.N. Julie, S.P. Strand, K.M. Varum, K.I. Draget, C.T. Nordgard, Chitosan: gels and interfacial properties. *Polym.* **7**, 552–579 (2015)

# Perceptions of Adolescents on Hazards of Using Electronic Gadgets



D. Jyothi and E. Manjuvani

**Abstract** The main thrust of this study was to compare the adolescent boys' and girls' perception on hazards of using electronic gadgets. The study was conducted on 100 boys and 100 girls aged 13–15 years. The sample was selected randomly from Students of 8th, 9th, and 10th standards studying in higher secondary schools at Tirupati. The data was collected by using structured questionnaire on perception of hazards of electronic gadgets. The results indicated that majority (56%) of adolescents had medium level of perception whereas 40% had low perception on hazards of using electronic gadgets. The adolescent boys had significantly higher perception on hazards of using electronic gadgets compared to their female counterparts.

**Keywords** Hazards · Electronic gadgets · Adolescents

## 1 Introduction

The usage of technology became an integral part of our daily life. From the last two decades there is a rapid utilization of electronic gadgets namely mobile phones, tablets, video games, television, and computers etc. Based on the technology more gadgets are being introduced with various new features and people are more dependent on it. On one side technology usage has positive role in our life, but on the other side it has many negative impacts as well [1].

One of the most existing environmental factors that influence adolescent's behavior is "electronic gadget". In recent years, electronic gadgets play an important role from school age children to old age all the people use electronic devices to have a better communication, research, employment etc. Currently educational institutions updated with digitalization, the students are using laptops, tablets and computers instead of books. The impact of electronic gadget on youth has grown steadily and it

---

D. Jyothi (✉)

Teaching Faculty, Sri Padmavathamma Govt. College of Nursing, Tirupati, India  
e-mail: [jyothidasaraju@gmail.com](mailto:jyothidasaraju@gmail.com)

E. Manjuvani

Department of Home Science, SPMVV, Tirupati, India

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_59](https://doi.org/10.1007/978-3-030-46939-9_59)

655

results in numerous physical, mental and social issues where they don't spend much time with people and activities around them. The physical health problems such as increased abdominal fat, insulin resistance, obesity, diabetes mellitus, sleep disturbances, musculoskeletal problems, visual and hearing problems. Due to spending more hours in using computers, it results in sleeping, increased living standards and reduced bed time [2]. The mental health problems namely mood disorders, hyper anxious, more irritable, depression as well as suicides. The social issues such as loneliness, poor interpersonal relationship and inactive participation in social activities. Prolonged usage of the electronic gadgets causes the above problems and consequently worsening their quality of life.

In India, internet users reached up to 500 millions in 2018. Of the 281 million daily internet users, 182.9 million (62%) urban people utilize internet daily where as only 98 million users from rural area, about 143 million were females using internet i.e., approximately 30% of total internet users.

This study aims to compare adolescent boys and girls perception on hazards of using electronic gadgets.

## 2 Objectives

1. To assess the perception of adolescents on hazards of using electronic gadgets.
2. To compare the perception of adolescent boys' and girls' on hazards of using electronic gadgets.

## 3 Method

### 3.1 Sample

All the students studying eighth, ninth and tenth standards of three schools at Tirupati, Andhra Pradesh constituted the population. Simple random sampling was used to select 100 boys and 100 girls for the present study.

### 3.2 Tool

The perception of adolescents was assessed by using structured questionnaire on hazards of electronic gadgets usage. It consists of 38 statements under the headings of causes of health hazards and prevention of health hazards.

## 4 Results and Discussion

The usage of technology is the modernization and creativity of individual. The modernization and creativity of technology attracts the attention of human beings for healthy development of relationships. Unfortunately especially adolescents and adults are at risk for the prevailing health hazards. Health problems are more due to usage of electronic gadgets are bad posture, headache, physical fatigue, poor sleep, stress, and compromised immunity.

Table 1 shows the distribution of perceptions level of adolescent boys' and girls' on hazards of using electronic gadgets.

Smart phone became part of human life especially for adolescents. Use of social networking system instead of smart phone, adolescents can be protected from smart phone addiction in order to engage in using mobile messengers or gaming apps. It indicates that positive function using of social networking system can prevent smart phone addiction.

In the present study, 16% of adolescent boys and 24% of adolescent girls had low perception level on hazards of using electronic gadgets. Majority (31%) of adolescent boys had medium level of perception and 25% of adolescent girls. Only three percent and one percent of adolescent boys and girls had high level of perception on hazards of using electronic gadgets.

Siraj (2015) conducted study on usage of internet and academic performance among 186 university students in Malaysian University. The author found that most (73%) of the subjects were females and residing in campus (69%), 36 (20.5%) students were using internet dependently. On whole 64.4% respondents were aware that internet was the supplement to the content given by lecturer [3].

Subba and Mandelia (2013) conducted study on ringxiety and usage of mobile phone among 336 subjects (173 were males and 163 were females). He found that 34.5% of students experienced ringxiety. They concluded that the medical students felt inconvenience in using the mobile phone and suffered with ringxiety. However this issue needs to be identified by all stake holders to be aware on symptoms and take measures how to reduce it [4].

Table 2 shows mean and S.D values of boys and girls on component of hazards of using electronic gadgets.

**Table 1** Distribution of adolescent boys' and girls' perception levels on hazards of using electronic gadgets

Adolescents	Level of perception on hazards of using electronic gadgets						Total	
	Low		Medium		High		N	%
	n	%	n	%	n	%		
Boys	32	16	62	31	6	3	100	50
Girls	48	24	50	25	2	1	100	50
Total	80	40	112	56	8	4	200	100

**Table 2** Mean perception scores of adolescent boys' and girls' on hazards of using electronic gadgets

Perception on hazards of using electronic gadgets	Adolescents				t-test
	Boys		Girls		
	Mean	SD	Mean	SD	
Health hazards	30.2	5.61	17.67	3.51	20.07**
Prevention of hazards	9.04	1.42	4.51	1.89	19.29**
Total	39.26	5.57	22.18	4.19	24.29**

\*\*Significant at 0.01 level

The mean perception scores on health hazards of using electronic gadgets among adolescent boys and girls were 30.2 and 17.67 respectively, hence the mean perception scores of adolescent boys on hazards of electronic gadgets was higher than their female counterparts and the difference were significant at 0.01 level.

Sundus (2018) conducted study on impact of gadget usage among children aged 12 years in USA. He reported 29% of toddlers used gadgets and rest of them (70%) were masters in primary school age itself. He found that over usage of gadgets causes vision problems such as myopia when children spend time for more than 8 h per day on gadgets. Children may have both positive and negative impact on their health due to usage of gadgets. The positive aspect on children using gadgets can develop good motor, cognitive, compilation skills and may not have any risk of injury or threat compared to outside playing. Negative aspects regarding gadgets utilization were delayed speech because of lack of communication with human beings. Negative impact on character due to over use of gadgets were mentally intellectual than their actual age. He concluded that this is difficulty to keep away from the gadgets but leads to radical change in their daily life [5].

Yeon-jin Kim (2018) conducted study on effect of smartphone and internet over usage on depression and anxiety among 4854 Korean adults of age group 19–49 years of Catholic University. There were 2573 males (53.01%) and 2281 females (46.99%). The author found that there was relative risk factor for depression and anxiety from both heredity and optimal matching were 10% higher for smartphone addiction. He recommended that adults with over usage of smart phone may monitored closely related to mental health problems highlights the necessity to develop policies on prevention and management of smart phone addition [6].

Acharya (2013) conducted study on “common health effects of cell phones among college students of age group 17–23 years from urban and rural backgrounds of Hyderabad”. Only 112 (25.4%) students used mobile phones less than a year while the remaining 329(74.6%) students used it for over a year and health effects include headache (51.47%) and irritability anger (50.79%). Other similar psychological symptoms such as lack of concentration, insomnia, poor academic performance and anxiety etc. Physical health symptoms viz., body aches (32.19%), eye strain (36.51%) digital thumb (13.8%) noticed frequently. Author suggested that curtail time spent

on talking in smart phones helps to reduce heat or use head phones or loud speaker [7].

The adolescent boys' and girls' mean scores on perception of hazards prevention were 9.04 and 4.51 respectively. The adolescent girls had scored lower than adolescent boys and the difference were significant at 0.01 level.

In contrast to the above results Latha (2011) found that 124 of 200 samples were aware about adverse effects of using smart phones. Males (5%) and females (10%) felt that there was a need to reduce the unwanted effects [8].

Communication technologies have brought the revolutionary changes in the wireless system and are more responsible for the most of the diverting effects on the living beings [9]. He recommended some simple mobile and no one can prevent various hazards from radiation of mobile phones [9].

Dein (2013) found that "too much usage of electronic gadgets by adolescents can range from mild to severe hazards namely back ache, carpal tunnel syndrome, itchy eyes and sleeping problems which leads to lack of concentration which effect on their school performance". He recommended awareness campaign for parents, through correct utilization of media in order to avoid physical, behavioral and social hazards [10].

## 5 Conclusion

The adolescent boys had significantly higher perception on hazards of using electronic gadgets compared to their adolescent girls. The adolescents had medium perception level on hazards of using electronic gadgets. Hence they need much awareness on the prevention of future complications regarding excessive usage of electronic gadgets.

## References

1. S.M.M. Rana, Positive and negative effects of electronic gadgets to students. Daily news paper, dt: 3 Feb 2019
2. T.F. Dorofaeff, Sleep and adolescence. Do New Zealand teenagers got enough? *J. Pediatr. Child Health* **42**, 515–520 (2006)
3. H.H. Siraj, A. Salam, Internet usage and academic performance. *Int. Med. J.* **22**(2), 83–86 (2015)
4. S.H. Subba, C. Mandelia, Ringxiety and the mobile phone usage pattern among the students of a medical college in South India. *J. Clin. Diagn. Res.* **7**(2), 205–209 (2013)
5. M. Sundus, The impact of using gadgets on children. *J. Depress. Anxiety* **7**(1), 1–3 (2018)
6. Y.-j. Kim, Effects of internet and smart phone addictions on depression and anxiety. *J. Environ. Res. Public Health* **15**(5), 859 (2018)
7. J.P. Acharya, Common health effects of cell phones among college students. *J. Community Med. Health Educ.* **3**(1), 1–4 (2013)
8. R. Latha, Awareness of mobile phone hazards among Malaysian University students. *Health* **3**(7), 406–415 (2011)

9. K. Sukhdeep, Effects of mobile radiation and its prevention. *J. Comput. Sci. Mob. Comput.* **5**, 298–304 (2016)
10. N.A. Dein, Harmful effect of commonly used electronic devices on adolescence and its safeguard at Shebin El-Kom. *J. Nurs. Health Sci.* **2**(1), 32–46 (2013)

# Genomics in Big Data Bioinformatics



Tahmeena Fatima and S. Jyothi

**Abstract** In today's world where there is massive growth of genomic data, feature-selection methods are proved to reduce the complexity making analysis for disease detection much easier. There is huge research work carried in studying the genomes of various living beings. In such enormous data generation the technique of feature selection is focused on significantly reducing the complexity of genomic data which enables data sources to analyze the data and converting it into valuable information. This paper focuses on structured literature review of the feature-selection techniques which are currently used in analyzing the genomic data in big data analytics. There are diverse data sets covered in area of big data analytics such as hidden patterns, unknown correlations, and other insights by thorough examination of enormous data sets. It enables in developing an infrastructure that demonstrate various issues and exhibit challenges and addresses in developing a realistic method which proves to be efficient and effective approach which enables in identifying genetic variants for clinically individualized diagnosis and therapy.

**Keywords** Systematic review · Genomic big data · Feature selection

## 1 Introduction

The amount of genomic data is exponentially growing with a rapid rate with the advancement of computational methods [1]. Various complexity and veracity issues may rise in the medical field if data is used without proper pre-processing leading to storage, analysis, privacy and security challenges [2].

As genomic data is massive in volume, it is quiet complex to deal with such kind of data. Handling genomic data has become quiet complicated due to its certain features

---

T. Fatima (✉) · S. Jyothi

Department of Computer Scenice, Sri Padamavati Mahila Visvavidyalayam, Tirupati, India  
e-mail: [tahmi.fatima18@gmail.com](mailto:tahmi.fatima18@gmail.com)

S. Jyothi

e-mail: [Jyothi.spmvv@gmail.com](mailto:Jyothi.spmvv@gmail.com)

© Springer Nature Switzerland AG 2020

S. Jyothi et al. (eds.), *Advances in Computational and Bio-Engineering*,  
Learning and Analytics in Intelligent Systems 15,  
[https://doi.org/10.1007/978-3-030-46939-9\\_60](https://doi.org/10.1007/978-3-030-46939-9_60)

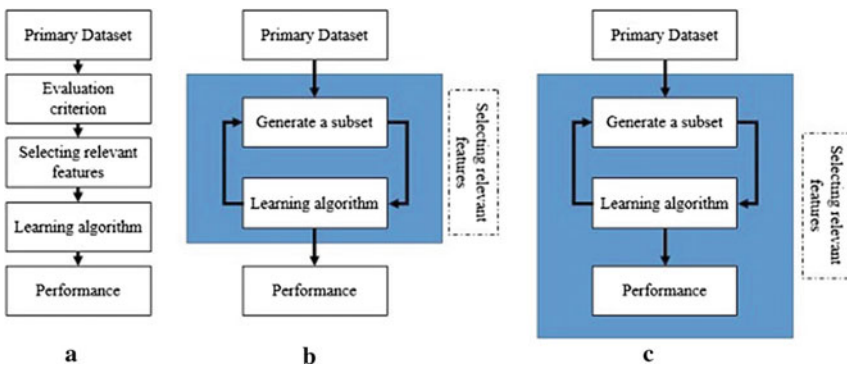


like complexity, heterogeneity and hybridity features. The process of dealing with such data is termed as Knowledge discovery process [3]. It includes various phases:

- (a) **Data recording:** This phase involves storing and capturing data including its tools and challenges.
- (b) **Data pre-processing:** In this step data is cleaned and preprocessed with the use of relevant tools that makes the data ready for analysis optimization.
- (c) **Data analysis:** In the step of data analysis different algorithms are used to evaluate data which is then followed by logical reasoning to this phase of data analysis includes evaluation of data using various algorithms followed with logical reasoning to give appropriate and correct meaningful outcomes.
- (d) **Data visualization and interpretation:** This phase of data visualization and interpretation appropriate methods for knowledge representation are used thereby providing the data relevancy and proximity.

The main objective of genomics is to discover and examine the out of sight causes of different genetic diseases by sequencing the genomes of various living beings. The edge of genetic sequencing has now shifted from sequencing to deploying already sequenced data. In the process of developing an appropriate model for preprocessing step one may come across multiple challenges while dealing with genomic data. These challenges can be overcome selecting the specific features and ignoring the irrelevant features which thereby decreases the volume complexity with the help of this feature selection method. The step of pre processing the data forms the base of accurate and precise investigation of genetically sequenced data. If the preprocessing step is done inaccurately it may lead to numerous challenges like features and attribute multiplicity, complexity in genomic data of small databases. Therefore, it is significantly critical to do the preprocessing step to achieve quality analysis and to overcome the above challenges [4].

Feature selection method enables the preprocessing step to achieve the goal of reducing the dimensionality and complexity of various dataset. Though many but there are three main types of feature selection methods.



- (i) **Filters:** Filter method is one of the preprocessing step which uses independent techniques to select various features which is independent of subsequent learning algorithms. A specific evaluation criteria is being followed to choose the set of features to assess the relevance and characteristic of each target variable to a certain degree [5].
- (ii) **Wrappers:** In the wrappers step of feature selection certain subset of characteristics are evaluated based on the precision of analytical model which are accomplished laterally. It uses classifier method for the evaluation of assumed subset of features and characteristics to examine its significance and relevancy. This method is not popular though efficient but computationally expensive [6].
- (iii) **Embedded:** It is a technique in which combines the qualities and features of filter and wrapper methods. Filter method are faster but not as efficient in comparison to wrapper method which are effective but computationally expensive whilst working with huge datasets. Apart from this there are other feature selection methods found in literature which are based on the above three basic types of methods.
- (iv) **Hybrid:** This method applies conjunctive features of the initial characteristics in feature selection methods successively [7].
- (v) **Ensemble:** This method as name resembles makes use of combined features and characteristics for the given subsets based on varied base classifiers. It incorporates the Use an aggregate of feature subsets of diverse base classifiers. It incorporates the practice of using diverse feature subsets [8].
- (vi) **Integrative:** It is one of the way that makes use of peripheral information which is gained to identify feature selection [9].

In the feature selection methodology of the big data processing cycle, specifically genomic data has gained much importance apparently in the upcoming days. There are various reviews given by many researchers based on survey analysis directed on methods to select the features of genome data and its role in enhancing the result quality. Vergara et al. [10] emphasizes and describes various issues that are addressed in the feature selection methods, by providing an implementation of mutual information feature selection framework. Li et al. [11] presents feature selection techniques and various classification methods in addition to the experimental implementations used in gene expression datasets. A new method of feature selection technique based on survey with its applications is presented by Wang et al. [12] which offers categorization of the gene feature selection in the area of bioinformatics applicable in big data analysis.

## 2 Big Data in Bioinformatics

With the advent of big data the volume of data in bioinformatics research is exponentially growing. The big data sources have vast information as they have exceeded the particle physics experiments and search—engine blogs and indexes. The data volume

is increasing in many field especially in bioinformatics research with the digitization of all processes and availability of high throughput devices at lower costs. For instance, the size of a single sequenced human genome is approximately 200 gigabytes [13]. The impact of rise in big data technologies is supported by increasing data volumes and decreasing cost of computing and analytics throughput. Biologist majorly are dependent on massive and continuously growing genomic data which are explored by different research groups to discover a novel biomarker rather than traditional laboratories. In the current trend bioinformatics in big data the technologies which capture the bio data has become more effective and cheaper like automated genome sequencers. There is massive growth of data in the field of bioinformatics in the recent years. In comparison from the year 2013 to 2014, the European Bioinformatics Institute (EBI), has gathered a huge increase of 40 petabytes of genetic data, proteins and related molecules in comparison to 18 petabytes, the largest biology-data repository [14]. As there is explosive growth with the increasing trend of genomic data the storage size of different organizations have multiplied in maintenance. The storage and processing power of various organizations are increasing with increasing massive bio-data store. The Hinxton data center cluster, installed by EBI does the maintenance of across 17,000 cores and 74 terabytes of RAM, to process their data. Apart from EBI many such organizations across the world are engaged in gathering and dispensing massive huge collection of biological and genome sequence data. Bioinformatics is very insightful field which generates highly voluminous data which in turn is very useful in accurate analytics in area of bioinformatics research. With the increase in huge bioinformatics data there may be rise in big data problems like particle physics data captured at CERN or high resolution satellite data received at NRSC/ISRO open data archive2 in comparison to the big data challenges. There are two major differences in such kind of data like most of the data generated are enormous and heterogeneous in nature. Heterogeneity is the most important aspect of bioinformatics data because most of the data generated are in need of numerous diverse and independent databases for interpretation and authentication.

Bioinformatics data are generated by various different firms which in turn are represented in diversified forms from variant sources consequently. The second major difference in bioinformatics is its massive and explosive growth of data in terms of dimensionality, multiple instances and its enormous distribution geographically over the globe. Various issues may arise such as inefficiency, cost and ethical issues due to failure of transferability of data because of the huge size because only few amounts of data can be transferred over the internet and rest of the data remains useless diminutive leading to inefficiency [15]. Henceforth, leading to various big data problems in the field of bioinformatics like geographical distribution apart from the challenges of dealing with the volume, velocity and variety. This in turn leads to remote sharing and analysis of data forcefully due to inefficient distribution of data geographically. There are many upcoming challenges in dealing with big data in bioinformatics. Therefore many policies of cloud computing technologies are used to tackle the problem of storage, processing and computation thereby overcoming these challenges by using the cloud for both [15]. Thus the cloud technology enables to overcome the challenges due to massive and huge amounts of data generated

remotely which are imposed by big data in the area of bioinformatics research. Beijing Genomics Institute (BGI), has launched Gaea which is a cloud-based analysis workflow system based on Hadoop frameworks which is one of the globally renowned center for sequencing of genomes. The Gaea installation by BGI has enabled genome analysis on an extensive basis which run across thousands of cloud-based computers parallel. Apart from Gaea, Bina provides another significant genome analytics solution which is based on cloud.

### 3 Types of Big Data in Bioinformatics

For various important research done in the network of various human diseases and gene related issues many different kinds of data are available. But basically five types of data are majorly used in bioinformatics research which are huge in size.

- (i) Data related to expression of genes,
- (ii) DNA data, RNA data, and protein sequence data,
- (iii) (PPI) protein-protein interaction data,
- (iv) Pathway data, and
- (v) (GO) gene ontology.

From different kinds of huge data sets gene expression analysis is done. In the analysis the levels of expression of innumerable genes are evaluated exposing it to various conditions like distinct developmental stages of various diseases and its treatment.

The best method of analyzing the gene expression level is based on microarray which is gene expression profiling. Microarray data are categorized into three kinds named gene-sample, gene-time, and gene-sample-time. In the profiles of gene expression there are differences of sample space record over time space in which sample space record the expression levels for varying external conditions, and in time space, expression levels are recorded based on different instances of time. Gene expression analysis uses the technique of comparing the expression values of infected and uninfected cells to analyze and identify the genes which are affected from pathogens or viruses. The outcome of the gene expression analysis help and assist the biomarkers in diagnosing various gene related diseases and also to identify its preventive measures. The result analysis of gene expression are stored and accessible from various sources of microarray data such as ArrayExpress4, Gene Expression Omnibus5 which are from EBI and NCBI respectively, and Stanford Microarray Database6. In sequence analysis, different analytical methods are applied to the DNA, RNA or peptide classifications are processed to understand their features, functions, structures, and evolution. There are various types of sequencing associated with DNA, RNA and sequence analysis methods.

## 4 Conclusion

**DNA sequencing** is the practice of defining the nucleic acid **sequence** in the study of genomes and proteins with its association of phenotypes. It includes various methods or technology in determining and identifying the potential drugs, identification of micro species present in an illustrative atmosphere, evolutionary biology, forensic identification etc. Sequence alignment and biological database search are done by using sequence analysis methodologies Apart from microarrays RNA sequencing is also used for various purposes like transmutation identification, credentials of post-transcriptional mechanisms, recognition of viruses and exogenous RNAs, and identification of Polyadenylation. In comparison of sequence analysis and microarray analysis sequence analysis is more effective than microarray analysis, because sequence data embed richer information. Sequence analysis makes use of sophisticated analytic tools and various computing infrastructures, while handling massive amount of sequence data generated in gene expression analysis [16].

## References

1. J. Andreu-Perez, C.C. Poon, R.D. Merrifield, S.T. Wong, G.Z. Yang, Big data for health. *IEEE J. Biomed. Health Inform.* **19**(4), 1193 (2015)
2. M. West, G.S. Ginsburg, A.T. Huang, J.R. Nevins, Embracing the complexity of genomic data for personalized medicine. *Genome Res.* **16**(5), 559 (2006)
3. C.P. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. *Inf. Sci.* **275**, 314 (2014)
4. D. Berrar, I. Bradbury, W. Dubitzky, Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics* **22**(10), 1245 (2006)
5. S. Landset, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *J. Big Data* **2**(1), 24 (2015)
6. N. Kushmerick, D.S. Weld, R. Doorenbos, *Wrapper Induction for Information Extraction* (University of Washington, Washington, 1997)
7. M. Naseriparsa, A.M. Bidgoli, T. Varae, A hybrid feature selection method to improve performance of a group of classification algorithms. arXiv preprint [arXiv:1403.2372](https://arxiv.org/abs/1403.2372) (2014)
8. A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategies for ensemble feature selection. *Inf. Fusion* **6**(1), 83 (2005)
9. B. Grasnack, C. Perscheid, M. Uflacker, A framework for the automatic combination and evaluation of gene selection methods, in *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pp. 166–174 (Springer, Berlin, 2018)
10. J.R. Vergara, P.A. Estévez, A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24**(1), 175 (2014)
11. T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* **20**(15), 2429 (2004)
12. L. Wang, Y. Wang, Q. Chang, Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods* **111**, 2 (2016)
13. R.J. Robison, How big is the human genome? *Precis. Med.* (2014)
14. EMBL—European Bioinformatics Institute, *EMBL-EBI Annual Scientific Report 2013* (2014)

15. V. Marx, Biology: the big challenges of big data. *Nature* **498**(7453), 255–260 (2013)
16. A. Nekrutenko, J. Taylor, Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* **13**(9), 667–672 (2012)