Algorithms for Intelligent Systems
Series Editors: Jagdish Chand Bansal · Kusum Deep · Atulya K. Nagar

Subhash Bhalla
Mangesh Bedekar
Rashmi Phalnikar
Sumedha Sirsikar *Editors*

Proceeding of International Conference on Computational Science and Applications

ICCSA 2021

Springer

Algorithms for Intelligent Systems

This book series publishes research on the analysis and development of algorithms for intelligent systems with their applications to various real world problems. It covers research related to autonomous agents, multi-agent systems, behavioral modeling, reinforcement learning, game theory, mechanism design, machine learning, meta-heuristic search, optimization, planning and scheduling, artificial neural networks, evolutionary computation, swarm intelligence and other algorithms for intelligent systems.

The book series includes recent advancements, modification and applications of the artificial neural networks, evolutionary computation, swarm intelligence, artificial immune systems, fuzzy system, autonomous and multi agent systems, machine learning and other intelligent systems related areas. The material will be beneficial for the graduate students, post-graduate students as well as the researchers who want a broader view of advances in algorithms for intelligent systems. The contents will also be useful to the researchers from other fields who have no knowledge of the power of intelligent systems, e.g. the researchers in the field of bioinformatics, biochemists, mechanical and chemical engineers, economists, musicians and medical practitioners.

The series publishes monographs, edited volumes, advanced textbooks and selected proceedings.

Indexed by zbMATH.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at https://link.springer.com/bookseries/16171

Subhash Bhalla · Mangesh Bedekar ·
Rashmi Phalnikar · Sumedha Sirsikar
Editors

Proceeding of International Conference on Computational Science and Applications

ICCSA 2021

Springer

Editors
Subhash Bhalla
School of Computer Systems
University of Aizu
Aizuwakamatsu, Japan

Rashmi Phalnikar
School of Computer Engineering
and Technology
Dr. Vishwanath Karad MIT World Peace
University
Pune, India

Mangesh Bedekar
School of Computer Engineering
and Technology
Dr. Vishwanath Karad MIT World Peace
University
Pune, India

Sumedha Sirsikar
School of Computer Engineering
and Technology
Dr. Vishwanath Karad MIT World Peace
University
Pune, India

ISSN 2524-7565 ISSN 2524-7573 (electronic)
Algorithms for Intelligent Systems
ISBN 978-981-19-0862-0 ISBN 978-981-19-0863-7 (eBook)
https://doi.org/10.1007/978-981-19-0863-7

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 Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Organizing Committee

Chief Patrons

Prof. Dr. Vishwanath Karad
Prof. Dr. Rahul Karad

Patrons

Dr. Ragunath Mashelkar
Dr. Vijay Bhatkar

Organizing Chair

Dr. Prasad Khandekar

Organizing Co-chair

Dr. Vrushali Kulkarni
Dr. Mangesh Bedekar

International Advisory Committee

Dr. Subhash Bhalla, Japan
Dr. Murli Vishwanathan, Australia
Dr. Andrew Stranieri, Australia

Dr. Xin-Wen Wu, Australia
Dr. Jay Gore, USA
Dr. Suresh Borkar, USA
Dr. Maode Ma, Singapore

TPC Chair

Dr. Rashmi Phalnikar

TPC Co-chair

Dr. Sumedha Sirsikar

Publication Chair

Dr. Aninda Bose

Technical Review Committee

Dr. Prasad Kulkarni, Professor, Electrical Engineering and Computer Science, University of Kansas
Dr. S. D. Joshi, Dean and Professor, Computer Department, BVCOE, Pune
Dr. Sanjeev Wagh, Professor and Head, Information Technology Department, COE, Karad
Dr. B. B. Meshram, Professor, Computer Engineering and Information Technology, VJTI, Mumbai
Dr. Bharat Chaudhari, Associate Dean, MIT-WPU, Pune
Dr.Anil Hiwale, Associate Dean, MIT-WPU, Pune
Dr. Balaji Patil, Associate HoS, School of CET, MIT-WPU, Pune
Dr. Bharati Dixit, Associate HoS, School of CET, MIT-WPU, Pune
Dr. Kailas Patil, Professor, Vishwakarma University, Pune
Dr. Shweta Dharmadhikari, PICT, Pune
Dr. Sachin Sakhare, Professor and Head, VIIT, Pune

Preface

The 3rd Springer International Conference on Computational Science and Applications (ICCSA 2021) was successfully organized by School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, during December 10–11, 2021. The objective of hosting ICCSA 2021 was to bring together experts for sharing of knowledge, expertise and experience in the emerging trends of computer engineering and sciences.

The conference highlighted the role of computational science in an increasingly interconnected world. The conference focused on recent developments in: scalable scientific algorithms, advanced software tools, computational grids and novel application areas. These innovations will drive efficient application in allied areas. The conference discussed new issues and new methods to tackle complex problems and identified advanced solutions to shape new trends.

Research submissions in various advanced technology areas were received and after a rigorous peer-review process with the help of program committee members and external reviewers, 25 papers were accepted. All the papers are published in Springer AIS series.

The conference featured paper presentations and expert talks on new technologies that were conducted by experts from academia and industry. Many distinguished personalities namely Dr. Subhash Bhalla, University of Aizu, Japan, Shri. P. M. Kurulkar, Director, Research and Development Establishment (Engrs.), Pune, Dr. Maitreya Natu, TCS, Pune, Manish Lunge and Ashwattha Sahastrabuddhe from Capgemini, Dinakaran Nianaygamurthy from IBM and Mr. Atulya Grover enlightened the participants.

Our sincere thanks to all special session chairs, distinguished guests and reviewers for their judicious technical support. Thanks to dynamic team members for organizing the event in a smooth manner. We are indebted to Dr. Vishwanath Karad MIT World Peace University for hosting the conference in their campus. Our entire organizing committee, faculty members of MIT-WPU and student volunteers deserve a mention for their dedicated efforts to make the event a grand success.

Special thanks to our program chairs for carrying out an immaculate job. We would like to extend gratitude to our publication chairs who did a great job in making the conference widely visible.

Lastly, our heartfelt thanks to all authors without whom the conference would never have happened. Their technical contributions to make our proceedings rich are praiseworthy. We sincerely expect readers will find the chapters very useful and interesting.

Aizuwakamatsu, Japan	Dr. Subhash Bhalla
Pune, India	Dr. Mangesh Bedekar
Pune, India	Dr. Rashmi Phalnikar
Pune, India	Dr. Sumedha Sirsikar

Contents

Part III Communication and Networking

Part IV Machine Intelligence

About the Editors

Dr. Subhash Bhalla joined the faculty of Jawaharlal Nehru University (JNU), New Delhi, in 1986, at the School of Computer and Systems Sciences. He was a postdoctoral fellow at Sloan School of Management, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA (1987–88). He is a member of the Computer Society of IEEE and SIGMOD of ACM. He is with the Department of Computer Software at the University of Aizu. He has also toured and lectured at many industries for conducting feasibility studies and for adoption of modern techniques. He has received several grants for research projects. Professor Bhalla currently participates in research activities on new query languages, big data repositories in science and astronomy, standardized electronic health records, polystore data management, edge computing and cloud-based databases. He is exploring database designs to support models for information interchange through the World Wide Web.

Dr. Mangesh Bedekar received his Ph.D. in Computer Science and Engineering from BITS Pilani, India. Prior to the Ph.D. program, he graduated from SSGMCE, Shegaon, and completed his master's at BITS Pilani, India. His primary research interests include Web data mining, Web personalization, user interface design, user interface improvements, browser customization, affective computing and information visualization. He has authored numerous national and international publications.

Dr. Rashmi Phalnikar has completed her Ph.D. in Computer Engineering from SVNIT Surat, India. Her research interests include data science and analysis, user interface design and software requirement engineering. She is currently guiding research scholars in these domains. She has published numerous research papers in renowned international journals and conferences.

Dr. Sumedha Sirsikar received Ph.D. in Computer Science and Engineering from SGBA University, MH, India. She graduated from Government College of Engineering Aurangabad and completed master's at Government College of Engineering Pune, MH, India. Currently, she is working as an Associate Professor at School of Computer Engineering and Technology, MIT-WPU, Pune. She has published more than 50 research papers in various renowned journals and International Conferences. Her research interests include Computer Networks and Security, Wireless Sensor Networks and IoT.

Part I
Computational Intelligence

Chapter 1
Novel Approach for Feature Selection Using Genetic Algorithm

Tanmay Unhale and Shilpa Sonawani

1 Introduction

When creating a predictive model, we need to reduce the number of features by the process of feature selection. The number of features needs to be reduced so that we lower the computational cost and, in some situations, also increase the model's performance. Hence, it is necessary to reduce the number of features while creating a predictive model. Statistical-based feature selection approaches involve applying statistical formulas to assess each feature's relationship with the target and choose the features that have the strongest relationship with the target. Although the choice of statistical measures for both the features and target variables is dependent on the data type, these methods can be quick.

1.1 Statistical-Based Methods

Variance threshold is a simple method for feature selection where it drops all features whose variance does not meet some threshold as it is assumed that features with a higher variance are most likely to contain more useful information. Features having a variance lower than or equal to the threshold value will be dropped. After dropping these features, we calculate the new dataset's accuracy and compare it with the original dataset. The formula for calculating variance is shown in Fig. 1.

T. Unhale (✉) · S. Sonawani
School of Computer Engineering and Technology, MIT World Peace University, Pune, India

S. Sonawani
e-mail: shilpa.sonawani@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_1

Fig. 1 Formula for calculating variance

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$$

Fig. 2 Formula for calculating Pearson correlation

$$r = \frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Pearson's correlation coefficient checks for the statistical relationship between two variables. It provides information on the direction and magnitude of the correlation. Correlations can be divided into two categories: positive and negative correlations. In a positive correlation, when feature M increases, feature N increases as well, and when feature M decreases, feature N decreases as well. Both features move in lockstep and have a straight relationship. In a negative correlation, if feature M increases, then there will be a decrease in feature N and vice versa. If two or more independent features are highly correlated, then they can be considered as duplicate features and one of them can be dropped. When these independent variables or features are highly correlated, then the change in one feature causes a change in the others, causing the model results to change. With a minute change in the data, the model outcomes tend to be unstable and fluctuate a lot. The formula for calculating Pearson's correlation coefficient is shown in .

The **F-score (or F1-score)** is another way to measure the accuracy of a model on a given dataset and is mostly used to evaluate binary classification systems. The F-score, mathematically, is defined as the harmonic mean of the model's recall and precision, which combines the model's precision and recall. Features with high F-score values are of importance. Figure 3 shows the formula for calculating the F1-score.

Fig. 3 Formula for calculating F1-score

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

2 Related Work

Hussein et al. [1] developed a GA which used selection methods such as elitism. The recognition engine is their own nearest neighbor-type classifier, which targets the optimization of the GA. The set of feature weights corresponds to the chromosome as a whole. For feature weighting runs, each weight is in the range of [0, 1]. More accurate classifiers are produced from the GA-based feature importance in the 200 sample/phase trials (87%). It accomplishes this and also reduces the feature count by around 5%. However, this improvement in accuracy and feature reduction increase the processing time by around 100–150%. Anusha and Sathiaseelan [3] devised an algorithm for optimal clustering that employs three goal functions to increase inter-cluster distance (diversity) while minimizing intra-cluster distance (compactness) with excellent accuracy. They offer an improved feature selection approach for multi-objective optimization problems using the k-means evolutionary algorithm in this study. This method removes outliers from the cluster set and produces objects that are very near to the desired characteristic. By maximizing cluster variety and accuracy while limiting cluster compactness, the suggested technique simultaneously maximizes three objectives. The suggested feature selection method uses NAGA II, and NLMOGA is used to evaluate the algorithm's performance and efficiency. The algorithm given by Blessie and Karthikeyan [5] is to apply the t-test to choose the most significant features from the dataset. The t-value is used to determine the significance of a t-distribution-based correlation coefficient. The feature is significant and is picked if the t-value is bigger than the crucial value at the 0.05 significant level. Apart from F-score, ROC, and accuracy, Sokolova et al. [6] recommend evaluating classifier performance using criteria other than F-score, ROC, and accuracy. For feature selection, Munson and Caruana [7] used bagging and bias–variance. Bagging is a type of meta-learning approach in which sub-samples are created at the time of model training for each individual sample. Feature selection helps in identifying the different features that deliver the best balance between the bias of having less number of features and the variance of having large number of features, according to these trials. As a result, the feature inclusion/exclusion threshold is determined by the learning algorithm.

3 Genetic Algorithm

The genetic algorithm (GA) is an evolutionary algorithm based on Charles Darwin's natural selection theory, which supports the survival of the fittest. The fittest individuals are chosen to create offspring, according to natural selection theory. Cross-over and mutation are used to pass on the features of the fittest parents to their offspring, ensuring that they have a greater chance of survival. This algorithm is a randomized search algorithm that generates high-quality optimization solutions by imitating biologically inspired natural selection processes such as selection, cross-over, and mutation.

3.1 Terminology for Genetic Algorithm

For the stochastic search process to begin, a **population** has been generated which contains a set of optimal solutions. Then, the GA will repeat over numerous generations until it discovers a solution that is both acceptable and optimal. This is the first generation that is created at random. One candidate solution in a generation or population is represented by a **chromosome**. A genotype is another name for a chromosome. Genes that carry the value for the optimal variables make up a chromosome. The **fitness function**, also known as the objective function, examines each individual solution in each generation in order to determine which individuals are the fittest.

3.2 Different Genetic Operators

Selection is the process of choosing the best solution from a population and having that solution acts as the parents of the next generation of solutions. This allows the strong characteristics to be passed down to the following generation. Selection can be done using ranked selection based on the fitness value. Then, genes from the two fittest parents are randomly exchanged to generate a new genotype or solution, which is known as **cross-over** or recombination. Based on the parent's segments of DNA transferred, a cross-over might be a one-point cross-over or a multi-point cross-over.

After selection and cross-over have generated a new population, it is randomly transformed by mutation. **Mutation** is a means of randomly changing a genotype in order to increase population variety and find better and more optimal solutions.

3.3 Function Definition

Here are some of the functions which are used in this algorithm. initialization_of_population(): This function is used to generate a random population for the stochastic search process that contains a set of possible solutions. While creating this population, 30% of the features are randomly dropped from every chromosome. This population is further sent to the objective function to calculate the fitness score.

fitness_score(): This function calculates the model accuracy for every chromosome in the population and sorts them in descending order, with the best accuracy at the top. It then returns the list for selection.

selection(): In the selection function, we select the best n_parents (depending upon the value given) who have had the best accuracy score up till now and they are sent for cross-over recombination.

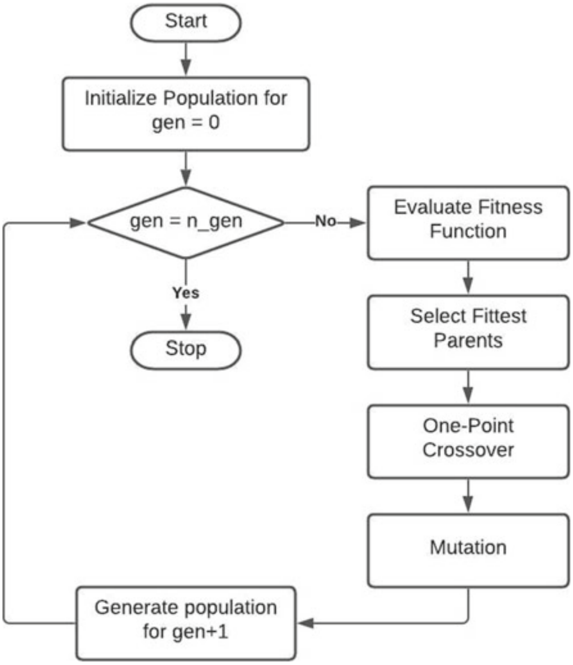
crossover(): In the cross-over function, the first half of the genes is picked from one parent, while the second half comes from the other parent. The goal here is to exchange genes from the fittest parents in order to create a new genotype or solution. mutation(): After cross-over function, the chromosomes undergo mutation in which the chromosome is randomly modified (depending on the mutation_rate) in order to promote diversity in the population and find an optimized solution.

generation(): This is the main function in which all the parameters such as n_parents, mutation_rate, and n_gen are given. Here, all the above functions are executed for the specified number of generations (n_gen) and the 5 best solutions (chromosomes) along with the prediction score are returned.

3.4 *Working of the Proposed Genetic Algorithm*

Figure 4 shows the algorithm that is proposed in this paper. Initially, a random population is generated which can be considered as generation = 0. In the population, there are a number of chromosomes having multiple genes where each gene is equivalent to a feature. Out of these, 30% of the features are randomly dropped.

Fig. 4 Proposed algorithm


Start
Initialize Population for gen = 0
gen = n_gen
No
Yes
Stop
Evaluate Fitness Function
Select Fittest Parents
One-Point Crossover
Mutation
Generate population for gen+1

Followed by that, calculations are performed in the fitness function where every chromosome is evaluated and the accuracy is determined for that set of features. Out of these, the fittest parents, i.e., the chromosomes having better accuracy, are sorted in descending order and are selected. After selection, one-point cross-over recombination takes place where the first and the second half of the genes are picked from the first and the second parent, respectively, to create new children. In the final stage, a random mutation takes place on these new children in which 20% of the genes are randomly altered. In the end, this mutated population is sent again as the population for the next generation and this entire procedure is repeated for ‘n’ generations.

3.5 Datasets Used

In the **Breast Cancer Diagnostics** [20] dataset, the features were taken after analyzing images of the breast mass that show the properties of the cell which are present in that image. From the given data, which consists of values like radius, texture, area, compactness, etc., we can diagnose whether the tumor is malignant or benign.

In the **Parkinson’s Disease** [21] dataset, features were computed by monitoring the effects of the disease over time, which required the patient to visit the clinic multiple times. Since most of the characteristic exhibited by a PD patient are vocal features, they were collected for diagnosis.

In the **Poly-cystic ovary syndrome** [22] dataset, the data was collected from patients from multiple hospitals across Kerala, India, and contains all the different parameters which are used to determine PCOS-related issues.

4 Experimental Result

In the first run, all three datasets are trained using all the features and accuracy is measured for seven classifiers. The best classifier for each dataset with the highest accuracy can be seen in Table 1. Initially, the best accuracy for the Breast Cancer, Parkinson’s Disease, and PCOS datasets is 97.2%, 91.8%, and 88.9%, respectively.

Table 2 shows the results of best accuracy after applying feature selection for different threshold values for variance. Features with variance lower than the threshold value are dropped, models are trained for a different number of features, and the best score out of all is selected. The highlighted values are the best accuracies. With this method, there was an improvement of 0.7% for the Breast Cancer dataset, 2% for the Parkinson’s Disease dataset, and 0.8% for the PCOS dataset. Here, the better results are given by different classifiers.

In Table 3, there are the results of the best accuracy after applying feature selection for different correlation values. Features that are highly correlated are dropped. Models are trained for a different number of features, and the best score out of all

Table 1 Score before applying any feature selection

	Breast cancer		Parkinson's disease		PCOS	
	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy
0	RandomForest	**0.972028**	LinearSVM	**0.918367**	RandomForest	**0.889706**
1	Logistic	0.965035	KNeighbors	0.897959	RadialSVM	0.860294
2	KNeighbors	0.965035	DecisionTree	0.897959	AdaBoost	0.860294
3	LinearSVM	0.958042	RandomForest	0.877551	LinearSVM	0.852941
4	GradientBoosting	0.958042	RadialSVM	0.877551	DecisionTree	0.845588
5	RadialSVM	0.951049	GradientBoosting	0.857143	GradientBoosting	0.838235
6	AdaBoost	0.951049	Logistic	0.836735	KNeighbors	0.698529
7	DecisionTree	0.93007	AdaBoost	0.836735	Logistic	0.676471

Table 2 Best accuracy after feature selection using variance threshold

	Breast cancer		Parkinson's disease		PCOS	
	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy
0	RandomForest	0.972028	LinearSVM	0.877551	RandomForest	0.860294
1	Logistic	0.951049	KNeighbors	0.836735	RadialSVM	0.698529
2	KNeighbors	0.972028	DecisionTree	0.897959	AdaBoost	0.852941
3	LinearSVM	**0.979021**	RandomForest	**0.938776**	LinearSVM	**0.897059**
4	GradientBoosting	0.958042	RadialSVM	0.877551	DecisionTree	0.860294
5	RadialSVM	0.951049	GradientBoosting	**0.938776**	GradientBoosting	0.852941
6	AdaBoost	0.965035	Logistic	0.836735	KNeighbors	0.676471
7	DecisionTree	0.972028	AdaBoost	0.918367	Logistic	0.867647

Table 3 Best accuracy after feature selection using Pearson correlation

	Breast cancer		Parkinson's disease		PCOS	
	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy
0	RandomForest	0.965035	LinearSVM	0.877551	RandomForest	0.875
1	Logistic	0.944056	KNeighbors	0.857143	RadialSVM	0.698529
2	KNeighbors	0.965035	DecisionTree	0.897959	AdaBoost	0.860294
3	LinearSVM	**0.972028**	RandomForest	**0.938776**	LinearSVM	**0.904412**
4	GradientBoosting	0.958042	RadialSVM	0.877551	DecisionTree	0.860294
5	RadialSVM	0.944056	GradientBoosting	0.836735	GradientBoosting	0.823529
6	AdaBoost	0.951049	Logistic	0.897959	KNeighbors	0.705882
7	DecisionTree	**0.972028**	AdaBoost	0.918367	Logistic	0.897059

Table 4 Best accuracy after feature selection using F-score

	Breast cancer		Parkinson's disease		PCOS	
	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy
0	RandomForest	**0.979021**	LinearSVM	0.877551	RandomForest	0.867647
1	Logistic	0.951049	KNeighbors	**0.918367**	RadialSVM	0.867647
2	KNeighbors	**0.979021**	DecisionTree	0.836735	AdaBoost	0.889706
3	LinearSVM	**0.979021**	RandomForest	0.877551	LinearSVM	**0.904412**
4	GradientBoosting	0.965035	RadialSVM	**0.918367**	DecisionTree	0.727941
5	RadialSVM	0.951049	GradientBoosting	0.897959	GradientBoosting	0.698529
6	AdaBoost	0.965035	Logistic	0.836735	KNeighbors	0.889706
7	DecisionTree	**0.979021**	AdaBoost	**0.918367**	Logistic	0.889706

these is selected. The highlighted values are the best accuracies. With this method, there was an improvement of 0% for the Breast Cancer dataset, 2% for the Parkinson's Disease dataset, and 1.5% for the PCOS dataset.

Table 4 shows the results of the best accuracy after applying feature selection using F-score. Features with a high F1-score are taken and tested with different permutations and combinations. The highlighted values are the best accuracies. With this method, there was an improvement of 0.7% for the Breast Cancer dataset, 0% for the Parkinson's Disease dataset, and 1.5% for the PCOS dataset.

As we can see, these statistical methods are limited and could only show an improvement of 0-2%. In order to improve the accuracy, the same dataset is tested with the GA.

In Table 5, we can see the best score in every generation for which the algorithm was run. Different classifiers were tested on all the datasets, but random forest and decision tree gave the best results when used in the fitness function for evaluation. With this method, there was an improvement of 2% for the Breast Cancer dataset, 6% for the Parkinson's Disease dataset, and 3% for the PCOS dataset.

Table 5 Best accuracy given by genetic algorithm

Breast cancer		Parkinson's disease		PCOS	
Random forest		Decision tree		Random forest	
Generation	Accuracy	Generation	Accuracy	Generation	Accuracy
1	0.986013	1	0.938775	1	0.904411
2	0.986013	2	**0.979591**	2	0.904411
3	0.986013	3	0.959183	3	0.904411
4	**0.993006**	4	0.959183	4	0.904411
5	**0.993006**	5	**0.979591**	5	**0.919117**

Table 6 Final comparison table

	Accuracy comparison				
	Original	Variance threshold	Pearson correlation	F-score	Genetic algorithm
Breast cancer	0.972028	0.979021	0.972028	0.979021	0.993006
Parkinson's disease	0.918367	0.938776	0.938776	0.918367	0.979591
PCOS	0.889706	0.897059	0.904412	0.904412	0.919117

Table 6 is the final comparison table showing the best accuracy given by every method for feature selection. Here, we can see a significant improvement in accuracy, showing the effectiveness of this algorithm.

5 Conclusion

After performing various tests by changing the threshold values and using different classifiers for statistical methods and different values for population, mutation, and number of generations for GA, a lot of data was generated. Out of these, the best accuracy is taken for comparing the different feature selection methods. Here, we can see that the results given by the GA are significantly better than the statistical-based methods for feature selection, showing an improvement of 2–6% accuracy and thus demonstrating the power of this algorithm, along with proving that randomized search algorithms like GA can give great results. The Breast Cancer, Parkinson's Disease, and PCOS datasets have been used for testing and have given satisfying results.

References

1. Hussein F, Kharma N, Ward R. Genetic algorithms for feature selection and weighting, a review and study. In: Proceedings of sixth international conference on document analysis and recognition. Available: https://doi.org/10.1109/icdar.2001.953980
2. Lingaraj H (2016) A Study on Genetic Algorithm and its Applications. Int J Comput Sci Eng 4:139–143. Available: https://www.researchgate.net/publication/309770246_A_Study_on_Genetic_Algorithm_and_its_Applications
3. Anusha M, Sathiaseelan J (2015) Feature selection using K-means genetic algorithm for multi-objective optimization. Proc Comput Sci 57:1074–1080. Available: https://doi.org/10.1016/j.procs.2015.07.387
4. Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond Accuracy, F-Score and ROC: a Family of Discriminant Measures for Performance Evaluation. In: Sattar A, Kang B (eds) AI 2006: Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol 4304. Springer, Berlin, Heidelberg. Available: https://doi.org/10.1007/11941439_114

5. Kumar S, Chong I (2018) Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. Int J Environ Res Public Health 15(12):2907. Available: https://doi.org/10.3390/ijerph15122907
6. Blessie E, Karthikeyan E (2012) Sigmis: a feature selection algorithm using correlation based method. J Algorithms Comput Technol 6(3):385–394. Available: https://doi.org/10.1260/1748-3018.6.3.385
7. Munson MA, Caruana R (2009) On Feature Selection, Bias-Variance, and Bagging. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science, vol 5782. Springer, Berlin, Heidelberg. Available: https://doi.org/10.1007/978-3-642-04174-7_10
8. Chaikla N, Qi Y (1999) Genetic algorithms in feature selection. 5:538–540, vol.5. 10.1109/ICSMC.1999.815609. Available: https://www.researchgate.net/publication/3828816_Genetic_algorithms_in_feature_selection
9. Fung G, Liu J, Lau R (1999) Signature verification based on a fuzzy genetic algorithm. 121–148. Available: https://www.researchgate.net/publication/262218121_Signature_verification_based_on_a_fuzzy_genetic_algorithm
10. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning (1st. ed.). Addison-Wesley Longman Publishing Co., Inc., USA. Available: https://doi.org/10.5555/534133
11. Kelly J, Davis L (1991) A hybrid genetic algorithm for classification. In Proceedings of the 12th international joint conference on Artificial intelligence—Volume 2 (IJCAI'91). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 645–650. Available: https://doi.org/10.5555/1631552.1631558
12. Razali NM, Geraghty J (2011) Genetic Algorithm Performance with Different Selection Strategies in Solving TSP 2. Available: https://www.researchgate.net/publication/236179245_Genetic_Algorithm_Performance_with_Different_Selection_Strategies_in_Solving_TSP
13. Tsai C-F, Eberle W, Chu C-Y (2013) Genetic algorithms in feature and instance selection. Know.-Based Syst 39:240–247. Available: https://doi.org/10.1016/j.knosys.2012.11.005
14. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J. Mach. Learn. Res 3, null (3/1/2003), 1157–1182. Available: https://doi.org/10.5555/944919.944968
15. Van der Putten P, van Someren M (2004) A Bias-Variance Analysis of a Real-World Learning Problem: The CoIL Challenge 2000. Mach Learn 57:177–195. Available: https://doi.org/10.1023/B:MACH.0000035476.95130.99
16. Haindl M, Somol P, Ververidis D, Kotropoulos C (2006) Feature Selection Based on Mutual Correlation. In: Martínez-Trinidad J.F., Carrasco Ochoa J.A., Kittler J. (eds) Progress in Pattern Recognition, Image Analysis and Applications. CIARP 2006. Lecture Notes in Computer Science, vol 4225. Springer, Berlin, Heidelberg. Available: https://doi.org/10.1007/11892755_59
17. Hall MA, Smith LA (1999) Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference. AAAI Press 235–239. Available: https://doi.org/10.5555/646812.707499
18. Demsar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res 7:1–30. Available: https://www.researchgate.net/publication/220320196_Statistical_Comparisons_of_Classifiers_over_Multiple_Data_Sets
19. Khandelwal R (2021) Genetic algorithm optimization algorithm. Medium (Online). Available: https://pub.towardsai.net/genetic-algorithm-optimization-algorithm-f22234015113
20. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, Archive.ics.uci.edu (2021) (Online). Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
21. UCI Machine Learning Repository: Parkinsons Data Set. Archive.ics.uci.edu, 2021 (Online). Available: https://archive.ics.uci.edu/ml/datasets/Parkinsons
22. Polycystic ovary syndrome (PCOS). Kaggle.com, 2021 (Online). Available: https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos

Chapter 2
A User-Driven COVID-19 Diseases Contact Tracing System

Pranav Ghadge, Piyush Pokharkar, Aprajita Jain, and Trupti Baraskar

1 Introduction

COVID-19 has caused a lot of loss of life and is continuing to do so, and a lot of citizens believe that the virus is here to stay. Citizens must learn to live with it. The COVID-19 is a highly contagious disease with an above-average mortality rate; living with it sounds like a difficult task. This calls for some solution on how to live with a deadly virus among us. COVID-19 prevention norms set by the World Health Organization say that you should maintain a social distance of at least 6ft, masks to be always worn, and not touch the face. But sometimes, it is violated. To keep the norms to check, government has come up with a few technological solutions, i.e., "contact tracing mobile applications." These kinds of apps help the user in knowing whether they met a COVID-19 positive patient or not in the past 15 days. This majorly serves for effective contact tracing and warns the user about the infection he/she may get due to the contract.

The available applications proposed by the government have few drawbacks. These applications ask for a lot of data, and users are reluctant about it. The major objective of this proposed work is to provide data security to users and simply implement contact tracing applications.

In this proposed work, hardware is implemented that can be utilized by all age groups of citizens, irrespective of their access to a smartphone. The used hardware does not depend on any databases, Internet, or platform. The hardware does not follow the traditional contact tracing approach but involves the user's awareness and intervention.

The user should check his/her COVID grade which is calculated based on the questions and answers. Further, it understands the specifics of the symptoms that

P. Ghadge (✉) · P. Pokharkar · A. Jain · T. Baraskar
School of Computer Engineering and Technology, MIT World Peace University, Pune, India
e-mail: trupti.baraskar@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_2

patients have. In this entire process, the answers of the user and his encounters are shared, but their identities are not revealed. During the setup of the hardware itself, each hardware is assigned a public key that acts as the user ID. It needs to be generated only once, whereas the answers to the questions can be updated and encounter logs can be downloaded anytime when the user required them. The objective of this proposed work is to provide a self-awareness, platform-independent, and generic contact tracing hardware to the masses that is economically viable and portable. This would improve the check on the spread of the infection and be useful in other applications where the knowledge of encounters and their specific information when broadcasted will prove to be beneficial.

2 Related Work

This COVID-19 contact trace app development paper [1–5] deals with the call for technological and location-based solutions in response to the pandemic, which allows for contact tracing efforts to be automated. Since the sole reason for the uncontrolled increase in the number of cases is the contact with the infected patient. So that contact tracing is very commencement of this pandemic has proved to be essential. The served paper [6, 7] has done work about Bluetooth smartphone apps that deal with the central idea of maximizing the reach to almost 90% of the population by using a predictive model, Bayesian model with already existing technological solutions under development. It identifies the stages in contact tracing and suggests different implementations for effective contact tracing applications. These papers [8, 9] study is limited to effective implementation and not covering actual usability is the gap identified. This paper [10] addressed the major concerns and doubts that arise in mind when contact tracing is invoked. It proves that digital solutions can be effective to prevent a potential outbreak if they are addressed properly, and the stakeholders' impact is reviewed. The paper [11] identifies those digital solutions do work when made mandatory and the privacy concerns that come along with them are addressed. Also, it shows that the citizens need to be assured that long-term adherence to it will not be necessary if there is effective contact tracing. The paper [12] suggests an idea that has been hypothetically devised called the hybrid bidirectional contact tracing which involves manual as well as technological contact tracing. This paper [13] identifies the possible causes of the failure of the contact tracing applications in the USA which cover the identification of social, political ramifications, and economical causes that were some potential reasons for the failure. This paper [14] was the aspect of assuming that the citizens were around the mob who actively are asymptotic carriers, and the fear of contact tracing apps was restricted to the specific privacy and social concerns. This paper [15] deals with the central idea of maximizing the reach to almost 90% of the population by using a predictive model, Bayesian model with already existing technological solutions under development. It identifies the stages in contact tracing and suggests different implementations for effective contact tracing.

3 Implementation

This proposed work has used two modules for implementation. The hardware module involves, ESP-32 (using the Espressif IoT development framework), and the software module involves, HTML page, serverless webpage (HTML/JavaScript).

3.1 Hardware Module ESP32

The ESP32 has an inbuilt client and server. The server provides outside connectivity to the ESP32 device by the user for configuration and scanning of the device. The ability to set the ESP32's BLE identification with a unique ID and health-code number is referred to as "configuration" (of the user). The ESP32 constantly advertises its ID and data. "Scan" in the hardware refers to retrieving any health-related connections which might have been sent. The BLE identification scan looks for the identifiers of other BLE devices in the vicinity. The cryptographic "#A99:" appears in any and all contact tracing identifiers used by every system. Any identities that do not begin with this letter will be disregarded. Those who begin with it are registered further into the internal RAM of the ESP32. The following are the hardware and software requirements for the development of the prototype.

- Desktop/laptop to run the webpage.
- Chrome web browser (version 83+).
- Bluetooth module: ESP 32 WROOM.
- Specifications: Low power consumption and generic Wi-Fi + BT + BLE MCU module.
- Portable power source (2A 5 V output capacity).

3.2 Software HTML Page-Serverless Webpage

Chrome's web-Bluetooth API can be used. It allows the user to connect with BLE devices and retrieve data.

•Bluetooth API: JavaScript
•GATT services: C.

3.3 Block Diagram

Below block diagram has three entities, i.e., user, hardware, and webpage. These are sufficient for the development of contact tracing applications (Fig. 1).

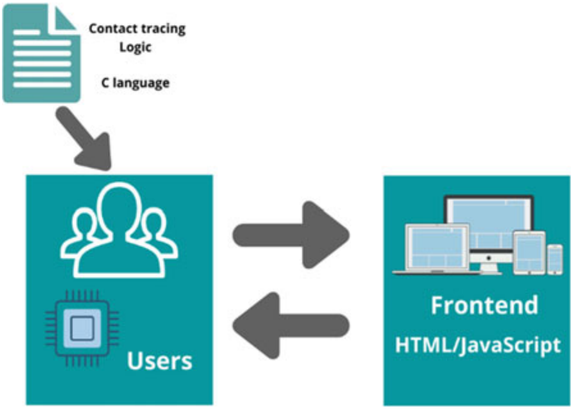


Fig. 1 Block diagram

3.4 The Schematic Representation of Implemented Application

A schematic representation of the components of the proposed application and their interactions is depicted in Fig. 2.

The processing of the implemented application is mainly in two phases. Firstly, broadcast its own information, i.e., ID and grade, with user responses, and secondly,

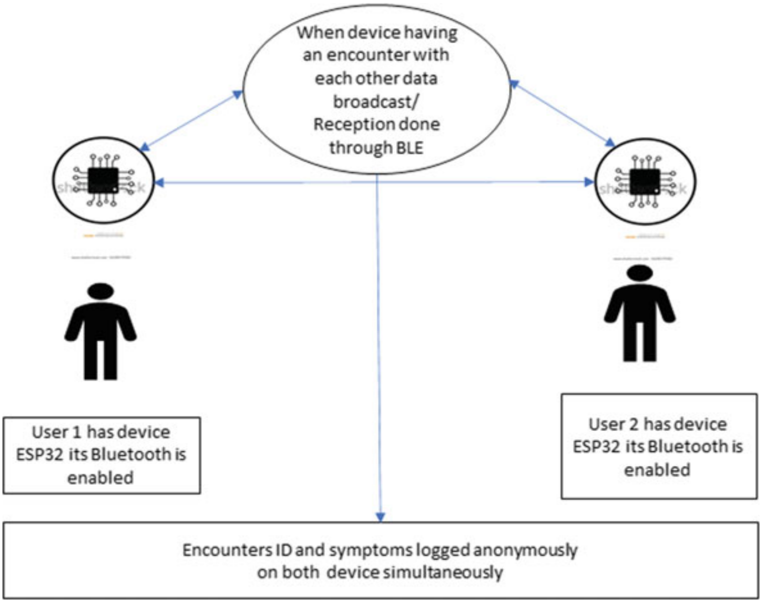

When device having an encounter with each other data broadcast/ Reception done through BLE
User 1 has device ESP32 its Bluetooth is enabled
User 2 has device ESP32 its Bluetooth is enabled
Encounters ID and symptoms logged anonymously on both device simultaneously

Fig. 2 Schematic representation of the components

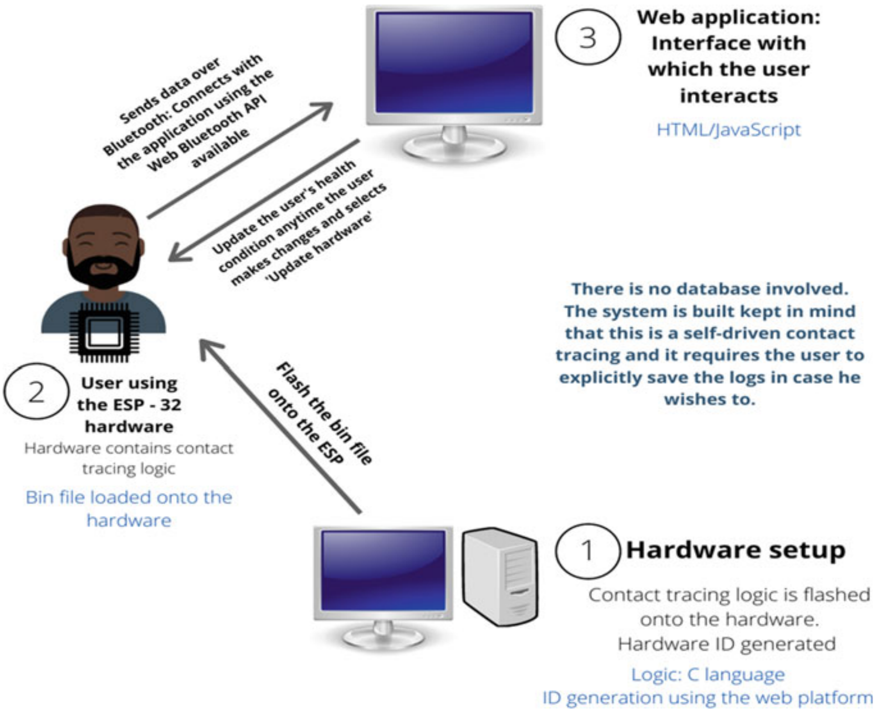

Web application: Interface with which the user interacts
3
HTML/JavaScript
Sends data over Bluetooth: Connects with the application using the Web Bluetooth API available
Update the user's health condition anytime the user makes changes and selects 'Update hardware'
There is no database involved. The system is built kept in mind that this is a self-driven contact tracing and it requires the user to explicitly save the logs in case he wishes to.
2
User using the ESP - 32 hardware
Hardware contains contact tracing logic
Bin file loaded onto the hardware
Flash the bin file onto the ESP
1
Hardware setup
Contact tracing logic is flashed onto the hardware.
Hardware ID generated
Logic: C language
ID generation using the web platform

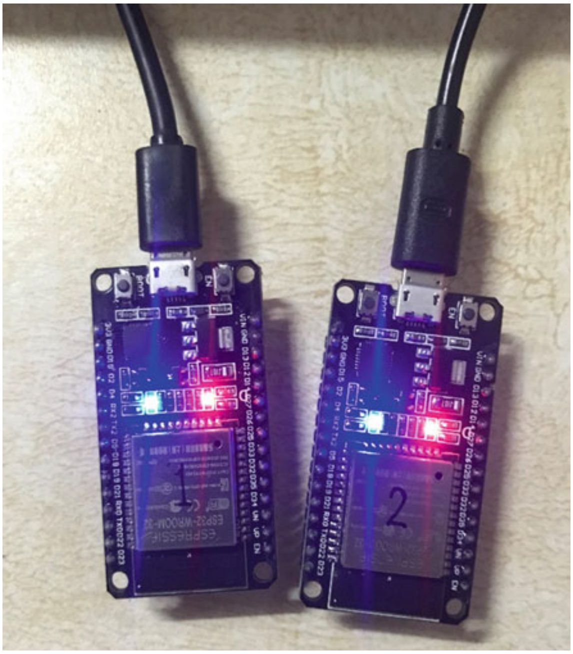
Fig. 3 Process of implemented contract tracing algorithm

log the same information from people around the user. The contact tracing starts with configuring the ESP32 in order to make it ready for broadcasting or logging information. Figure 3 shows the whole process of implemented contract tracing algorithm.

This configuration needs to be done only once before use. Following are the steps used in configuring the hardware.

(1) Initially, the hardware is connected to any desktop/laptop, and a flashing utility is downloaded for the hardware to be flashed with the logic.
(2) The bin file to be loaded onto the hardware, and the non-server-based web application file is also downloaded.
(3) Now, using the flashing software, the bin file is flashed onto the hardware.
(4) Next, open the Html file on the browser and connect the ESP via the Chrome web-Bluetooth API available.
(5) In case, a problem persists in connecting to the Bluetooth, manually inspect the ESP.
(6) Once connected to the web application, generate the unique ID for the ESP using the generate ID button. This ID is unique and is calculated using the MD5 algorithm.

Fig. 4 Contract tracing between two devices



(7) There is another round of MD5 to generate a public code that is added with a prefix #A99 and then latter 4-digit hex-health conditions to validate the entries that the hardware is logging and prevent any entry that is out of the scope of the ESP. The name of the ESP is set using the esp_ble_gap_set_device_name() function. For storing each contact, 24 hex-digits are needed.

(8) After generating the ID, mark the relevant health conditions and click on the update hardware button to update the information on the hardware. This would reflect in the change in the 4-digit hex-values depending upon the user's choice of health responses. The hardware is ready to be used.

Figure 4 Contact tracing between two ESP32 devices.
Figure 5 Setup used during implementation.

4 Contact Tracing Algorithm

After configuring the hardware, it can be used efficiently to perform the self-driven contact tracing. When the ESP is being used in daily life and encounters another device that has an active Bluetooth Low Energy (BLE), then logging of the encounters is done in the following steps.

(1) When the hardware is connected to a power source, the ESP starts working and keeps the RED LED on as a sign to show it is functioning.

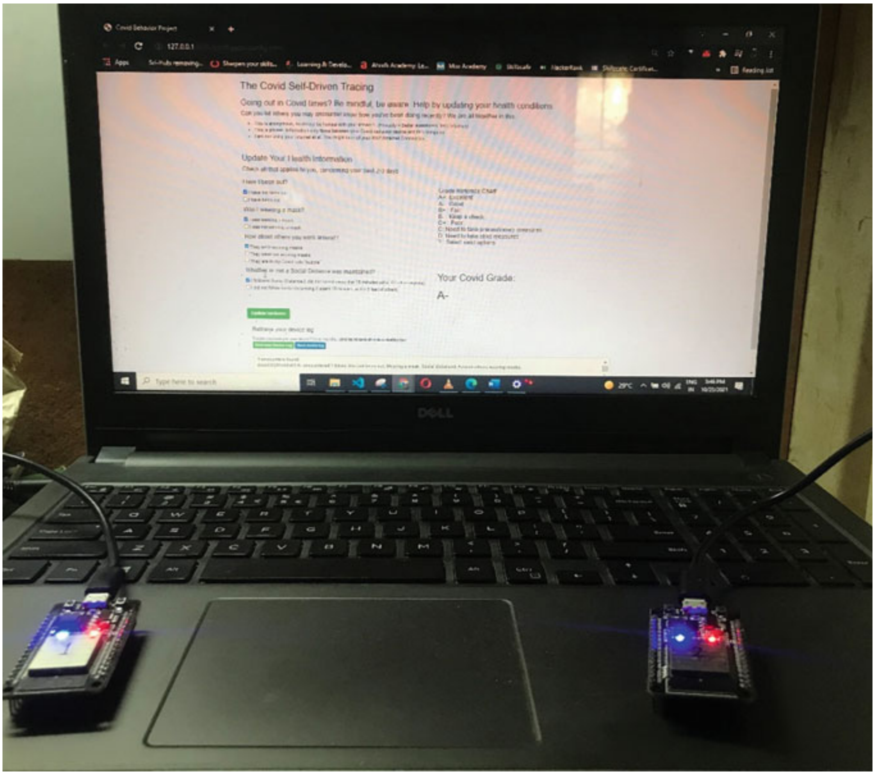


Fig. 5 Setup used while contract tracing between two devices

(2) On supplying power, the esp_ble_gap_start_advertising () function starts advertising the parameters to devices that have BLE enabled and stops it when the power is disconnected and the ESP_LOGI() method gets the GATTS parameters.

(3) The IDs of all the active BLE devices are logged into the ESP's memory using the ESP_GAP_BLE_SCAN_RESULT_EVT(). But not all entries are of the contact tracing hardware. Hence, the fake_test() method firstly validates the entry with the prefix #A99 and health condition hex-value and then logs the encounter into the ESP using the add_encounter() method.

(4) This logging involves validating the previous encounters or new encounters using the seen_ within_dt() or the make_last() functions.

(5) When the same encounter occurs within a time period of 10 min, the count is not updated, but if there lies an entry between the encounter of hardware or the time period of 10 min has elapsed the encounter count is incremented by the seen() function.

(6) When the ESP has filed the encounters, it shows a blue LED that signals that it has some logs to be downloaded.

(7) The ESP can later at any point be connected to the non-server-based webpage to read the logs from the ESP to be displayed to the user. The logs in the ESP are made using the example_write_event_env() function that uses the esp_ble_gatts_send_response() inbuilt function.

(8) Anytime, the ESP is erased the entire configuration process needs to be repeated before its use.

5 Result Analysis

The test cases were designed and tested based on the availability of two prototypes of contact tracing hardware devices, smartwatches in the surroundings, and Bluetooth-enabled speakers.

The concept of logging the unique ID number of devices and the information that they broadcast was tested here. The prototypes are programmed in such a way they broadcast their UID and COVID status grade. The grade is obtained by answering a set of predefined basic questions relating to the COVID norms established by health authorities around the world. An environment was set up where two such devices were tested right from setting up the hardware until logging encountering events of the other hardware. The two hardware devices showed logs appropriately based on the responses updated in the devices as expected. Also, the vicinity of other BLE devices did not hamper the logs. No logs other than those of the #A99 were seen in the textbox. The functionalities of saving the logs, updating the hardware with new information, and resetting the entire hardware in case the logging fails were also tested. The experimental setup gave the expected results in all the cases. The experiment was thus successfully conducted.

The following are the key findings of our prototype testing:

(1) The accuracy of the system in logging the encounter with contact tracing hardware has given high and favorable results.

(2) Any other encounters are not logged when the user downloads the device logs.

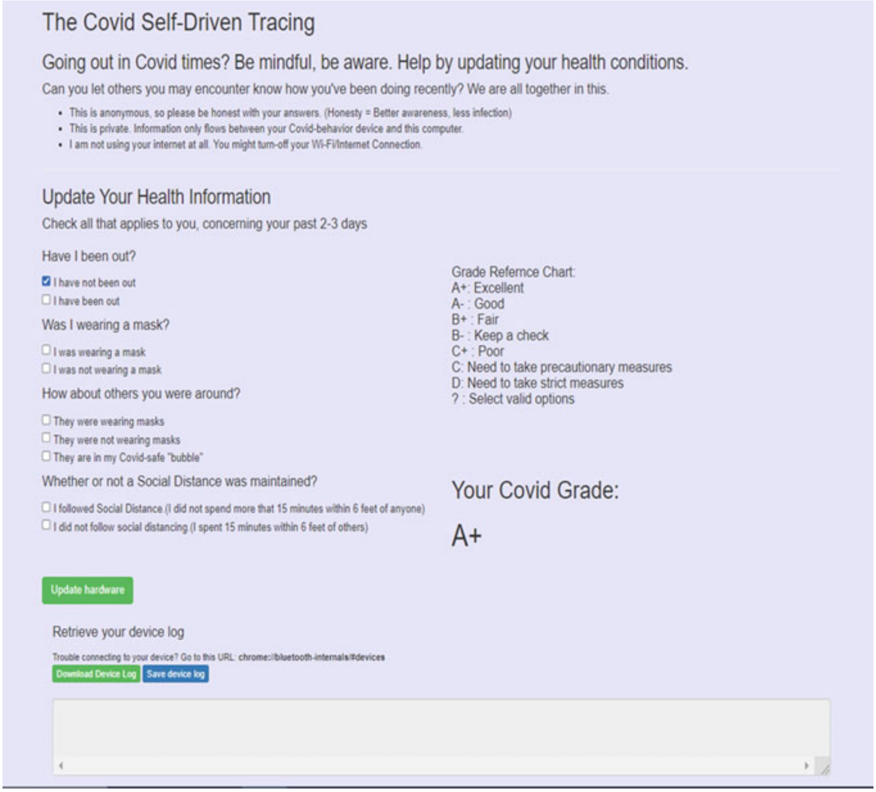
(3) The hardware readily updates when new information is updated.

(4) The prototype passed all the expected test cases thus, providing a piece of effective information to the user regarding his encounters.

Figures 6 and 7 show the first release application which is limited to only the COVID pandemic contact tracing as the questions are specific to its symptoms, precautions, and its grade is calculated based on the responses of the user. This grade is then advertised all the time to let the surrounding users aware of the level of precaution, and they need to adhere to at their personal levels.

Figures 8, 9, and 10 show various grade depending on the user entered data.

Figure 11 shows how the selection of the particular ESP32 is done via Bluetooth which then encounters the log.



The Covid Self-Driven Tracing

Going out in Covid times? Be mindful, be aware. Help by updating your health conditions.

Can you let others you may encounter know how you've been doing recently? We are all together in this.

- This is anonymous, so please be honest with your answers. (Honesty = Better awareness, less infection)
- This is private. Information only flows between your Covid-behavior device and this computer.
- I am not using your internet at all. You might turn-off your Wi-Fi/Internet Connection.

Update Your Health Information

Check all that applies to you, concerning your past 2-3 days

Have I been out?

☑ I have not been out
□ I have been out

Was I wearing a mask?

□ I was wearing a mask
□ I was not wearing a mask

How about others you were around?

□ They were wearing masks
□ They were not wearing masks
□ They are in my Covid-safe "bubble"

Whether or not a Social Distance was maintained?

□ I followed Social Distance.(I did not spend more that 15 minutes within 6 feet of anyone)
□ I did not follow social distancing.(I spent 15 minutes within 6 feet of others)

Grade Refernce Chart:
A+: Excellent
A- : Good
B+ : Fair
B- : Keep a check
C+ : Poor
C: Need to take precautionary measures
D: Need to take strict measures
? : Select valid options

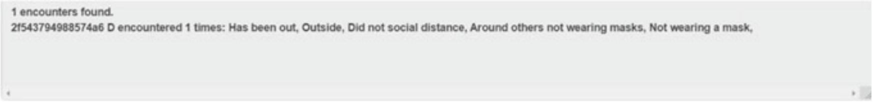
Your Covid Grade:

A+

Retrieve your device log

Trouble connecting to your device? Go to this URL: chrome://bluetooth-internals/#devices

Fig. 6 User registration page with grade calculation

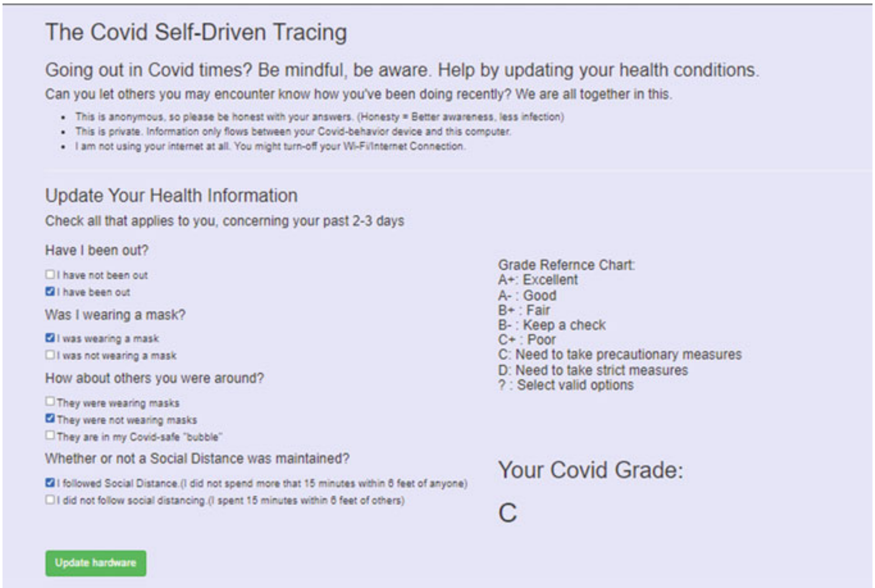


```
1 encounters found.
2f543794988574a6 D encountered 1 times: Has been out, Outside, Did not social distance, Around others not wearing masks, Not wearing a mask,
```

Fig. 7 Encounter log

6 Conclusion

This system aims to ensure effective contact tracing that would help the people stay safe and cautious of their contraction to the virus. Despite the advances in medical treatments and vaccines under testing, the major concern till the entire population is immunized is contracting the infection through an infected patient. The hardware developed is to expand the scope of contact tracing to all types of the population be it youth, middle-aged, or old age. The implemented system that can be improvised in the future for tracing diseases is not only limited to this pandemic, but other diseases that spread through contact and can be of major health concern. Also, its



The Covid Self-Driven Tracing

Going out in Covid times? Be mindful, be aware. Help by updating your health conditions.

Can you let others you may encounter know how you've been doing recently? We are all together in this.

- This is anonymous, so please be honest with your answers. (Honesty = Better awareness, less infection)
- This is private. Information only flows between your Covid-behavior device and this computer.
- I am not using your internet at all. You might turn-off your Wi-Fi/Internet Connection.

Update Your Health Information

Check all that applies to you, concerning your past 2-3 days

Have I been out?

☐ I have not been out
☑ I have been out

Was I wearing a mask?

☑ I was wearing a mask
☐ I was not wearing a mask

How about others you were around?

☐ They were wearing masks
☑ They were not wearing masks
☐ They are in my Covid-safe "bubble"

Whether or not a Social Distance was maintained?

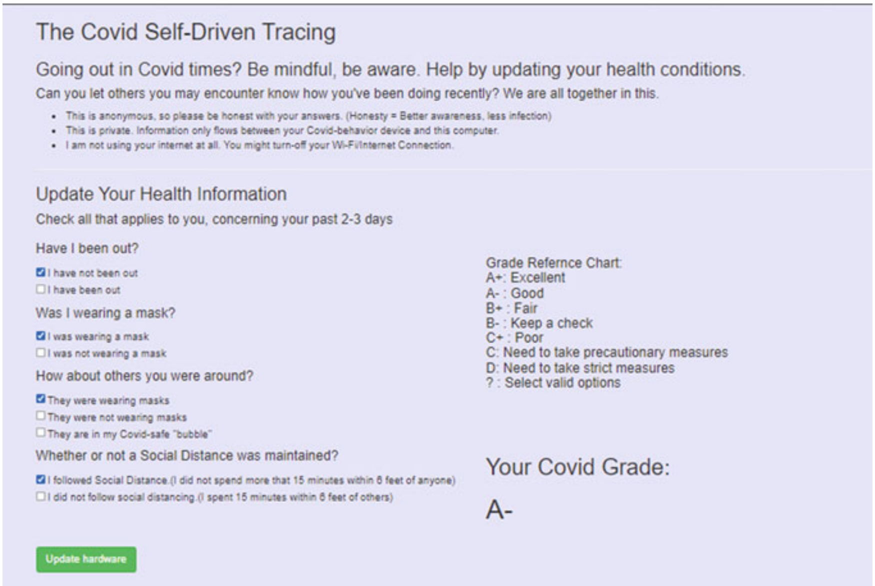
☑ I followed Social Distance.(I did not spend more that 15 minutes within 6 feet of anyone)
☐ I did not follow social distancing.(I spent 15 minutes within 6 feet of others)

Grade Refernce Chart:
A+: Excellent
A- : Good
B+ : Fair
B- : Keep a check
C+ : Poor
C: Need to take precautionary measures
D: Need to take strict measures
? : Select valid options

Your Covid Grade:

C

Fig. 8 Grade calculation 1



The Covid Self-Driven Tracing

Going out in Covid times? Be mindful, be aware. Help by updating your health conditions.

Can you let others you may encounter know how you've been doing recently? We are all together in this.

- This is anonymous, so please be honest with your answers. (Honesty = Better awareness, less infection)
- This is private. Information only flows between your Covid-behavior device and this computer.
- I am not using your internet at all. You might turn-off your Wi-Fi/Internet Connection.

Update Your Health Information

Check all that applies to you, concerning your past 2-3 days

Have I been out?

☑ I have not been out
☐ I have been out

Was I wearing a mask?

☑ I was wearing a mask
☐ I was not wearing a mask

How about others you were around?

☑ They were wearing masks
☐ They were not wearing masks
☐ They are in my Covid-safe "bubble"

Whether or not a Social Distance was maintained?

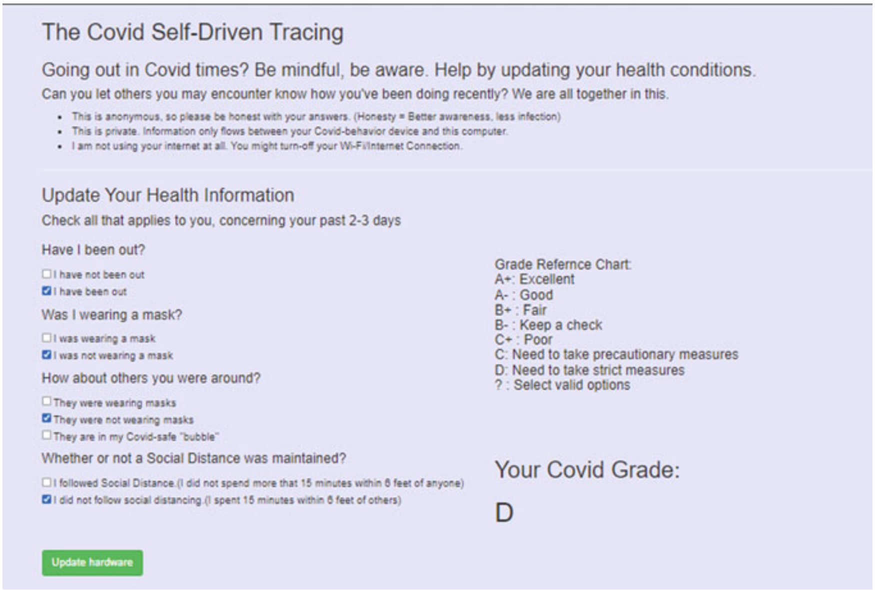
☑ I followed Social Distance.(I did not spend more that 15 minutes within 6 feet of anyone)
☐ I did not follow social distancing.(I spent 15 minutes within 6 feet of others)

Grade Refernce Chart:
A+: Excellent
A- : Good
B+ : Fair
B- : Keep a check
C+ : Poor
C: Need to take precautionary measures
D: Need to take strict measures
? : Select valid options

Your Covid Grade:

A-

Fig. 9 Grade calculation 2



The Covid Self-Driven Tracing

Going out in Covid times? Be mindful, be aware. Help by updating your health conditions.

Can you let others you may encounter know how you've been doing recently? We are all together in this.

- This is anonymous, so please be honest with your answers. (Honesty = Better awareness, less infection)
- This is private. Information only flows between your Covid-behavior device and this computer.
- I am not using your internet at all. You might turn-off your Wi-Fi/Internet Connection.

Update Your Health Information

Check all that applies to you, concerning your past 2-3 days

Have I been out?

☐ I have not been out
☑ I have been out

Was I wearing a mask?

☐ I was wearing a mask
☑ I was not wearing a mask

How about others you were around?

☐ They were wearing masks
☑ They were not wearing masks
☐ They are in my Covid-safe "bubble"

Whether or not a Social Distance was maintained?

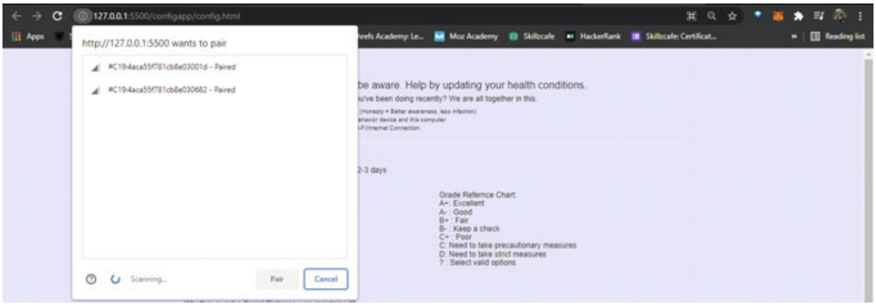
☐ I followed Social Distance.(I did not spend more that 15 minutes within 6 feet of anyone)
☑ I did not follow social distancing.(I spent 15 minutes within 6 feet of others)

Grade Refernce Chart:
A+: Excellent
A- : Good
B+ : Fair
B- : Keep a check
C+ : Poor
C: Need to take precautionary measures
D: Need to take strict measures
? : Select valid options

Your Covid Grade:

D

Fig. 10 Grade calculation 3



127.0.0.1:5500/configapp/config.html
Apps
Moz Academy
Skillzcafe
HackerRank
Skillzcafe: Certificat...
Reading list

http://127.0.0.1:5500 wants to pair

#C19:4aca55f781cb8e03001d - Paired

#C19:4aca55f781cb8e030682 - Paired

Scanning...

be aware. Help by updating your health conditions.

Grade Refernce Chart:
A+: Excellent
A- : Good
B+ : Fair
B- : Keep a check
C+ : Poor
C: Need to take precautionary measures
D: Need to take strict measures
? : Select valid options

Fig. 11 Encountering Bluetooth devices near by

scope can be expanded to smartphones to reduce dependability on the hardware. Basically, this application lies in self-driven contact tracing where the user in this pandemic situation can have a detailed view of the encounters throughout the day along with their symptoms answered through a specified set of questions that help in deciding the level of exposure to the virus and the steps to combat it. This proposed application is able to log the location module, and it can be reused for the hardware and smartphone. Also, the process of logging the encounter and maintaining a safe social distance are interlinked as the location gets logged only when the distance is violated and the user gets notified.

Acknowledgements The author thanks Dr. Vrushali Kulkarni HOS, School of Computer Engineering and Technology, MIT World Peace University, Pune, India, for providing all kinds of facilities and support.

References

1. Alex Berke MB (2020) Assessing disease exposure risk with location data: a proposal for cryptographic preservation of privacy. arXiv:2003.14412v2 [cs.CR]
2. Leith DJ (2020) Coronavirus contact tracing: evaluating the potential of using bluetooth received signal strength for proximity detection. arXiv:2006.06822 [eess.SP]
3. Sundaramoorthy K (2017) A novel approach for detecting and tracking humans. In: International conference on technical advancements in computers and communications
4. Michael K, Abbas R (2020) Behind COVID-19 contact trace apps: the google–apple partnership. In: IEEE Consumer electronics magazine, vol 9, issue 5, pp 71–76
5. Roba Abbas KM (2020) COVID-19 contact trace app deployments: learnings from Australia and Singapore. In: IEEE consumer electronics magazine
6. McLachlan S (2020) Bluetooth smartphone apps: are they the most private and effective solution for COVID-19 contact tracing? arXiv:2005.06621
7. Vaughan A (2020) There are many reasons why covid-19 contact tracing apps may not work. New Scientist. Last Accessed 9 May 2020. https://www.newscientist.com/article/2241041-their-are-many-reasons-why-covid-19-contact-tracing-apps-may-not-work/#ixzz6JtJYzqXw
8. Maccari L, Cagno V (2021) Do we need a contact tracing app? Comput Commun 166:9–18. ISSN 0140-664. https://doi.org/10.1016/j.comcom.2020.11.007. (https://www.sciencedirect.com/science/article/pii/S0140366420319873)
9. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler- Dörner L, Parker M, Bonsall D, Fraser C (2020) Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. Science
10. Schneider J, Love W, Rusie L, Flores A, Tadesse B, Hazra A, Munar D (2021) COVID-19 contact tracing conundrums: insights from the front lines. Am J Public Health 111:917–922. https://doi.org/10.2105/AJPH.2021.306200
11. Rodríguez P, Graña S, Alvarez-León EE et al (2021) A population-based controlled experiment assessing the epidemiological impact of digital contact tracing. Nat Commun 12:587. https://doi.org/10.1038/s41467-020-20817-6
12. Bradshaw WJ, Alley EC, Huggins JH et al (2021) Bidirectional contact tracing could dramatically improve COVID-19 control. Nat Commun 12:232. https://doi.org/10.1038/s41467-020-20325-7
13. Clark E, Chiao EY, Amirian ES (2021) Why contact tracing efforts have failed to curb coronavirus disease 2019 (COVID-19) transmission in much of the United States. Clin Infectious Diseases 72(9):e415–e419. https://doi.org/10.1093/cid/ciaa1155
14. Chan EY, Saqib NU (2021) Privacy concerns can explain unwillingness to download and use contact tracing apps when COVID-19 concerns are high. Comput Human Behav 119:106718. ISSN 0747-5632. https://doi.org/10.1016/j.chb.2021.106718. (https://www.sciencedirect.com/science/article/pii/S0747563221000406)
15. BlueTrace: a privacy-preserving protocol for community-driven contact tracing across borders, 9 Apr 2020 (Online). Available: https://bluetrace.io/static/bluetracewhitepaper938063656596c104632def383eb33b3c.pdf

Chapter 3
A Study of Genetic Mutations, Amplification, Deletion and Fusion in Endocrine Resistant Advanced Breast Cancer

Reena Lokare and Sunita Patil

1 Introduction

Breast cancer is one of the leading causes of death in countries like the US [1]. Among all cancer types, breast cancer is the second largest cancer seen in women. Recent advances in medical science have helped medical professionals in understanding the disease, detecting disease at an early stage and treating the disease effectively. These advances have reduced the number of deaths in women due to cancer. Breast cancer develops slowly and therefore is said to be more dangerous. Early diagnosis is much helpful in proper treatment of breast cancer. The chances of survival increase with early prognosis [2].

Recently, progress in understanding breast cancer has helped more accurate and effective treatment of breast cancer. Only clinical discussions and tests are not helpful in treating breast cancer patients. A personalized approach is required in selecting suitable treatment for a patient. Most of the time, breast cancer is detected in the third or fourth stage, and doctors get very less time for treatment [3]. Different treatment options are available to treat breast cancer. Treatment of breast cancer is said to be multidisciplinary as it includes locoregional (surgery and radiation therapy) and systemic therapy. Systemic therapies are endocrine therapy, for HR^+ breast cancer, chemotherapy, anti-HER2 therapy for HER2 positive patients, bone stabilizing agents, immunotherapy, etc. Various clinical observations are done like gender, age, stage of cancer, type of cancer: Primary/Metastatic, to name a few. Some treatments are given in combination or individually. In medical terms, they are stated as first-line, second-line, third-line and fourth-line of therapy. Initially, the first-line

R. Lokare · S. Patil (✉)
K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India
e-mail: spatil@somaiya.edu

R. Lokare
e-mail: reena.l@somaiya.edu

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_3

of therapy is given to a patient. If a patient doesn't recover, a second-line of treatment is given to the patient. Doctors study and analyze reasons for failure at the end of first-line therapy and accordingly decide second-line of therapy and so on [4–6].

Looking at tremendous diversity in the disease in terms of symptoms, therapy and diagnosis, some personalized approach is required in the treatment. Recent progress in the field of medicine facilitates molecular profiling of tumor samples. Techniques like next generation sequencing and microarray help in molecular profiling of tumor samples. Understanding disease at molecular level helps doctors in complete understanding of the disease. Molecular profiling provides multi omics details of tumor such as genomics, epigenomics, transcriptomics, proteomics, metagenomics, etc. [7, 8].

Proposed study aims at discovering biomarkers in ER^+ breast cancer which can be treated with endocrine therapy. Looking at complexity in breast cancer, personalized therapy, a recent treatment for cancer is highly recommended. A tumor profiling provides complete molecular details of the underlying cancer. By understanding such molecular details of the tumor, personalized treatment is provided to the patient [9]. Here, endocrine therapy suitability is being studied for breast cancer. This study is focused on clinical as well as genomic data of Estrogen Receptor-Positive patients. These patients have taken hormonal therapy which doesn't suit them and resulted in changes in genes. The genomic landscape provides curated genomic data. This data provides details of gene alterations such as altered genes, amplified genes, deleted genes, fusion genes, etc.

Machine learning algorithms are applied on treatment and clinical data in order to study the correlation between hormonal therapy failure and gene alterations. Once genetic changes are understood, selection of hormonal therapy can be done effectively. This study focuses on studying altered, amplified, fusion and deleted genes. The information gained from this study can be used in personalized therapy recommendation.

2 Breast Cancer

Breast cancer is a heterogeneous disease having different clinicopathologic characteristics and constitutes heterogeneous tumors and gives different responses to the same treatments. Breast cancer tumors are classified based on clinical features, molecular features, etc. [10, 11]. The aim of this study is to find out biomarkers in every type of breast cancer. Once biomarkers are identified, accurate treatment can be offered to the patient.

2.1 Breast Cancer Classification

This section discusses classification of breast cancer into different subtypes. Different aspects like altered genes, clinical features, ER and HER2 status can be used for classifying breast tumors into different categories. Doctors can make such classification breast tumors and decide an accurate treatment strategy for a patient.

2.1.1 Based on Heterogeneity in Structure

Breast tumors are classified into multiple categories based on their morphology and structural organization.

Common cancer type: Invasive Ductal Carcinoma, not otherwise specified (IDC NOS)-observed in 75% cases [11].

Genes altered in Invasive Ductal Carcinoma are shown in following Fig. 1:

It is observed that TP53, PIK3CA, GATA3, ERBB2, CCND1, FGF19, FGF4, FGF3, FGFR1, MAP3K1, PTEN, MYC, KMT2C, RYR2, USH2A, TG, BRIP1, COL22A1, UBR5, LAMB3, CDK12, AKT1, NF1, ARID1A, DCAF4L2, KMT2D, MAP2K4, MUC16, ESR1, PDE4DIP.

TP53, PIK3CA, GATA3, ERBB2, CCND1 are highly mutated genes among all.

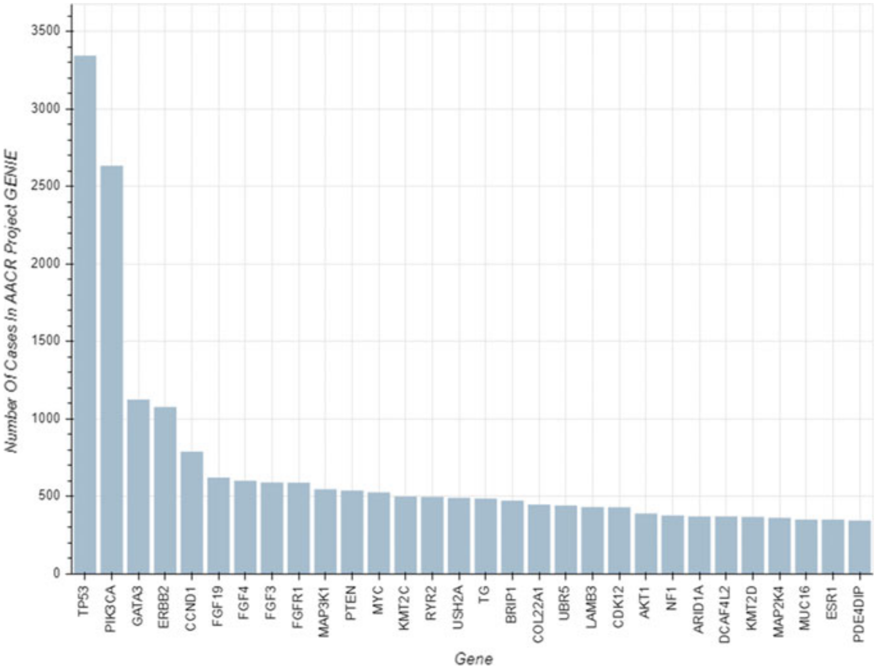


Fig. 1 Frequently altered genes in invasive ductal carcinoma [12]

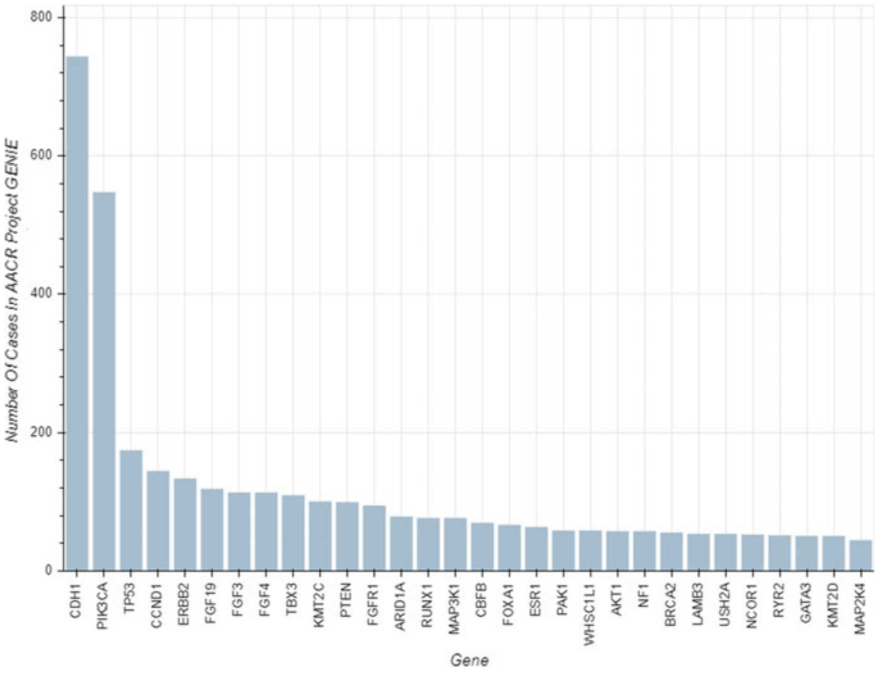


Fig. 2 Frequently altered genes in invasive lobular carcinoma

Most frequent histologic type breast cancer: Invasive Lobular Carcinoma (ILC) observed in 10% cases [11].

These two cancer types make about 90% of overall breast cancers.

Figure 2 shows the genes altered in Invasive Lobular Carcinoma.

It is observed that CDH1, PIK3CA, TP53, CCND1 and ERBB2 are more frequently altered in breast cancer [12].

2.1.2 Classification Based on Immunopathology

Main biomarkers identified in breast cancer are estrogen receptor (ER), progesterone receptor (PR) and human epidermal receptor 2 (HER2). ER markers indicate that tumors may respond to anti-estrogen on endocrine therapy. PR status does not indicate any clinical benefit from endocrine therapy. Monoclonal antibody trastuzumab treatment is given to HER2$^+$ cases.

Depending on specific markers identified in breast cancer, breast tumors are classified into following categories:

- ER$^+$ (ER$^+$/HER2$^-$): Genes mutated in ER$^+$ cancer is: PIK3CA, TP53, GATA3, MAP3K1, CDH1, MLL3, MAP2K4, PTEN, RUNX1, NCOR1, TBX3, AKT1, CTCF, NF1, PIK3R1, FAM47C, CBFB, SF3B1, TBL1XR1, ZFP36L1, FOXA1, TLR4, CDKN1B, GPS2, OR6A2, OR2L2, RB1, PTPN22 [13].

- HER2$^+$ (ER$^-$/HER2$^+$): Same genes mutated as in ER$^+$ [13].
- Triple negative (TN; ER$^-$/PR$^-$/HER2$^-$): Same genes mutated as in ER$^+$ [13].
- Triple positive (ER$^+$/PR$^+$/HER2$^+$): Genes mutated in triple positive cancer are ERBB2, PIK3CA, PIK3R2, AKT1, IGF1R, SMO, DOT1L, ERG and CEBPA [14].

2.1.3 Based on Intratumoral Heterogeneity

Breast cancer tumors can be classified based on the intratumoral heterogeneity. It exists at ER/HER2 status and also at genomic level. Genomic heterogeneity is found to be present in more than half of the tumor samples that were examined [11]. Intratumoral heterogeneity is one of the reasons for resistance to treatment and poor survival rate on overall cancer patients. Breast tumor shows variable threshold of positivity for the HER2 receptor in progesterone and estrogen receptors in breast tumors. Because of this variability, patients who take the same treatment for the same type of cancer get a variable response and hence result in failure of targeted therapies [14].

2.1.4 Molecular Heterogeneity Based on Gene Expression Level

Techniques like microarray, next generation sequencing, etc., facilitates molecular profiling of cancer types. Table 1 lists breast cancer subtypes: Luminal A and Luminal B. Luminal A subtype shows higher levels of ESR1, ER and ER-regulated genes whereas Luminal B shows decreased levels of the same genes. Also, proliferation is observed to be increased in Luminal A subtype and decreased in Luminal B subtype. So, Luminal A shows overall improved outcome but Luminal B shows relatively worse prognosis [11].

Table 1 Breast cancer subtypes

Molecular Subtype	Genes regulated by ER signaling pathway	Proliferation	Impact
Luminal A	Higher levels of ESR1, ER, ER-regulated genes	Decreased	Improved overall outcome
Luminal B	Decreased levels of ESR1, ER, ER-regulated genes	Increased	Relatively worse prognosis

2.1.5 Genomic Heterogeneity

Every breast cancer subtype shows genomic heterogeneity. Following genomic breast cancer subtypes are identified:

- Luminal A
- Luminal B (HER2)-
- Luminal B (HER2) +
- HER2 overexpression
- Basal
- Normal-like

Studies show that different genes are altered, amplified in all these breast cancer subtypes. However, some of the gene's alteration and amplification is found to be overlapped across different subtypes, termed as genomic heterogeneity [10, 11].

2.1.6 Microenvironmental Heterogeneity

Breast cancer has heterogeneity at microenvironmental level. Microenvironmental heterogeneity can be measured using image analysis of cancer and microenvironmental cells. It is found from this study that grade 3 tumors classified by microenvironmental grade 3 heterogeneity have 10-year disease-free (53%) survival rate than that of ordinary grade 3 tumors (70%). Mutation in TP53 is also found in the tumor samples which also have microenvironmental heterogeneity. So together microenvironmental heterogeneity and TP53 mutation can be used for more accurate prognosis and predicting disease-free survival of patients.

Such study of genomic alterations in tumor samples plays a role of significant biomarkers for offering personalized treatment to the patient [11, 16].

2.1.7 Macro Environmental Heterogeneity

Specific features are discovered in microenvironmental heterogeneity. Along with those features factors such as age, menopausal status, body mass index and immune status shall be considered. Along with study of gene alterations, considering such clinical factors give more accurate treatment decisions and hence increase in patient's disease-free survival months [11].

2.1.8 Longitudinal Heterogeneity

Longitudinal heterogeneity is nothing but the alterations in tumor features during progression. Looking at changes at genomic level, primary tumor and metastasis tumor have a lot of heterogeneity. This heterogeneity can be classified with respect to ER and HER2 status of tumor samples.

The biggest challenge in advanced breast cancers is dealing with spread of primary tumor, termed as metastasis tumor, changes happening on disease progression need to be studied and accordingly changes need to be done in treatment planned [11, 15].

3 Genomic Perspective of Breast Cancer

A study of advanced metastases breast tumors, which are treated with hormonal or endocrine therapy and there was a failure of therapy, is discussed here. The clinical, genomic and treatment data are studied here. In this study, clinical features such as age, menopausal status, detailed cancer type, overall tumor grade, nuclear grade, HR status, HER2 status, PR status, alive status, no of disease-free months, treatment duration, reason of stopping the treatment type of treatment given are available. Along with clinical and treatment data, genomic data is also available. This genomic data is available in various forms such as altered genes, amplified genes, fusion genes and deleted genes. Because of hormonal therapy, failure mutations and deletions are observed in the majority of the tumor samples. Also, these samples showed a shorter duration of response to hormonal therapies [17].

A detailed study of advanced metastases breast cancer with failure to endocrine/hormonal therapy:

i. Study of Altered Genes:

Altered genes are that genetic combination that do not occur naturally by mating or natural recombination. Such altered genes produce different proteins. Altered or mutated genes produce cancer causing tumors [18].

Here, a study of altered genes in metastases breast tumors is done. The aim of the study is to find details of genes altered before and after hormonal therapy. This study can give the information regarding altered genes, which can be used as a biomarker in personalized therapy recommendation.

Table 2 shows breast cancer subtypes in which alteration of genes have increased after giving hormonal therapy.

Figure 3 shows after giving hormonal therapy to HR+/HER2- breast cancer patients, number of gene alterations are shown to be increased in case of ESR1, CDH1, TP53, GATA3, PIK3CA, PTEN, ERBB2 and FGFR1 genes.

Figure 4 shows after giving hormonal therapy to HR-/HER2+ breast cancer patients, the number of gene alterations are shown to be increased in case of TP53 and ERBB2. This increase is too small i.e., 1%.

Figure 5 shows after giving hormonal therapy in HR+HER2+ breast cancer; there is no notable increase in gene mutation.

Figure 6 shows there isn’t a significant increase in gene mutation in case of triple negative breast cancer after hormonal therapy. Just TP53 gene has increased mutation by 1% of the samples after the therapy.

ii. Study of Amplified Genes:

Table 2 Frequency of gene alteration before and after hormonal therapy

	Breast cancer subtype							
Altered Gene	HR + /HER2-		HR-HER2 +		HR + HER2 +		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
ESR1	48	88	2	2	7	15	4	10
CDH1	104	132	6	7	14	18	9	13
TP53	208	290	7	21	29	40	21	32
GATA3	106	132	1	2	14	15	12	15
CCND1	102	166	5	4	12	25	15	19
PIK3CA	240	282	14	12	29	40	36	37
PTEN	63	72	1	3	3	6	3	14
ERBB2	77	114	5	11	4	20	6	19
FGFR1	77	109	5	2	8	10	10	11

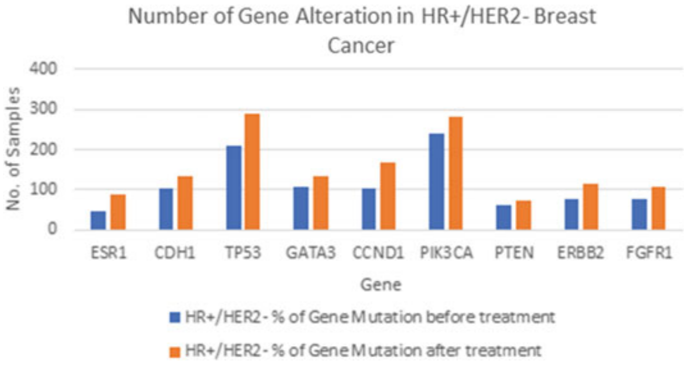


Fig. 3 Gene alteration in HR+ /HER2-Breast Cancer

Gene amplification is nothing but an increase in the number of copies of a particular gene. As breast cancer progresses, changes in environment, cell function may create copies of some genes. In this study, changes in gene alternation of amplified genes are done.

It is found that amplified genes are responsible for cancer progression in breast cancer. The frequently amplified genes are ERBB2, FGFR1, MYC, CCND1 and PIK3CA. Also, CCND2, EGFR, FGFR2, NOTCH3 are less frequently amplified genes [17, 19].

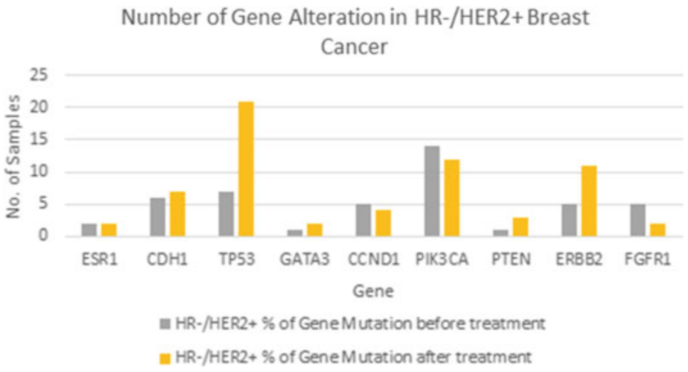


Fig. 4 Gene alteration in HR-/HER2 + in Breast Cancer

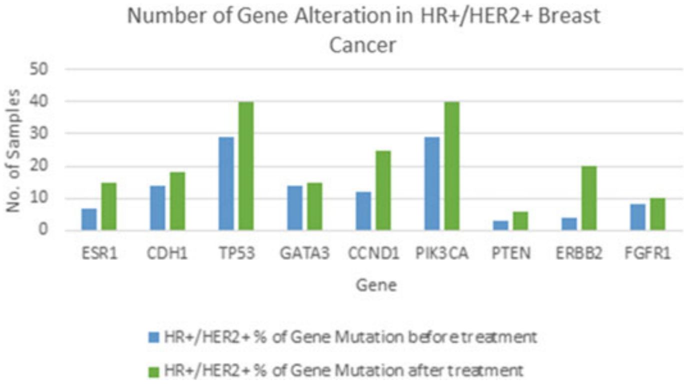


Fig. 5 Gene alteration in HR+/HER2+ Breast Tumor

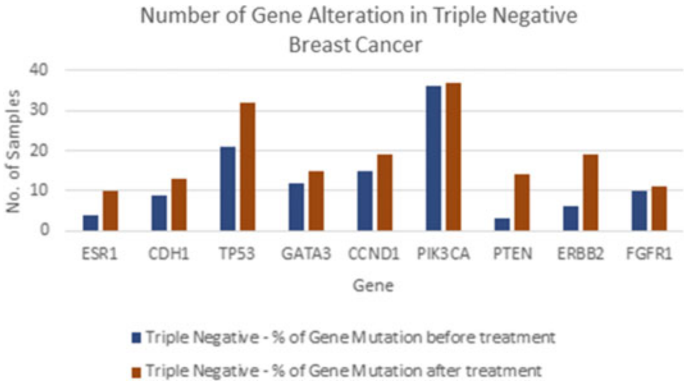


Fig. 6 Gene alteration in triple negative Breast Cancer

Table 3 lists the number of alternations in breast cancer subtypes: HR+/HER2-, HR-HER2+, HR+HER2+ and Triple negative.

Table 3 shows frequency of altered genes.

Figure 7 shows the changes in genes altered in amplified genes. This study shows that HR+/HER2-cancer has maximum gene alterations overall. Out of which, CCND1, ERBB2, SPOP, FOXA1, FGFR1 have more alterations compared to other genes. Also, these genes show increased alterations after hormonal therapy.

Figure 8 shows that HR-/HER2+ type breast cancer shows little increase in the number of mutations in ERBB2 gene. Overall, this cancer type has not much gene alterations.

Figure 9 shows that HR+/HER2+ type breast cancer shows an increase in the number of mutations in CCND1, ERBB2 and SPOP genes. Overall, this cancer type has not much gene alterations.

Figure 10 shows ERBB2, CCND1, FOXA1 and SPOP genes have increased mutations after hormonal therapy.

iii. Study of Deleted Genes:

Breast cancer shows deletion of specific genes on disease progression. [17]

Deletion of NF1:

NF1 gene is always mutated and deleted in many cancer types including breast cancer. Deletion of NF1 results in an increase in risk of developing cancer and also drug resistance. Deletion of NF1 has shown poor clinical outcome and is found in endocrine resistant advanced breast cancers [20].

Table 4 shows the frequency of gene alteration in deleted genes.

Figure 11 shows that there is significant increase in gene alteration in case of CCND1, ERBB2, SPOP, FOXA1 and FGFR2 genes in HR+/HER2- breast cancer. This particular cancer has no deletions of NF1 gene.

Figure 12 shows there aren't many mutations in studies of genes in HR-/HER2- breast cancer. However, it shows there is deletion of NF1 in 3 samples.

Figure 13 shows that there is significant increase in gene alteration in case of CCND1, ERBB2 and SPOP genes in HR+/HER2+ breast cancer. This particular cancer has no deletions of NF1 gene.

Figure 14 shows that there is significant increase in gene alteration in case of CCND1, ERBB2, SPOP, FOXA1 and FGFR2 genes in HR+/HER2- breast cancer. This particular cancer has no deletions of NF1 gene.

4 Study of Genes

The genes in cancer tumor samples play a vital role as biomarkers in personalized therapy recommendation to cancer, especially advanced cancer-like endocrine resistant metastases breast cancer studied here. A dataset of patients undergone hormonal therapy is studied. The study includes identifying change in gene mutations after hormonal therapy failure. It is found that due to hormonal therapy failure in advanced

Table 3 Frequency of gene alteration before and after hormonal therapy

	Breast cancer subtype							
Amplified Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
ESR1:AMP	6	15	1	0	0	1	0	2
CDK4:AMP	7	13	0	2	0	2	1	1
CDK6:AMP	2	4	0	0	1	1	0	0
MTOR: AMP	1	2	0	0	0	0	0	0
CDH1: AMP	3	7	0	0	0	0	0	0
TP53: AMP	3	4	0	0	0	1	1	0
GATA3:AMP	9	18	0	1	1	0	0	1
CCND1:AMP	101	163	5	4	12	25	15	19
PIK3CA:AMP	3	8	0	0	1	2	1	3
PTEN: AMP	13	19	0	1	1	1	1	2
ERBB2:AMP	52	85	3	7	2	15	4	11
ARID1A:AMP	1	1	0	0	1	0	0	0
ARID2:AMP	2	6	0	0	0	2	1	0
CREBBP:AP	3	9	0	0	0	0	0	0
NF1: AMP	4	7	3	0	0	0	1	1
FGFR4:AMP	2	2	0	1	0	1	1	0

(continued)

Table 3 (continued)

	Breast cancer subtype							
Amplified Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
SPOP: AMP	13	36	0	0	2	10	0	5
FOXA1:AMP	14	27	1	1	1	2	0	8
TBX3: AMP	7	11	0	0	0	3	0	3
CTCF: AMP	0	1	0	0	0	0	0	0
ERBB3:AMP	1	4	0	0	0	0	0	
KRAS: AMP	2	3	0	0	1	3	1	3
BRAF: AMP	3	1	0	0	0	0	0	0
MAP2K1:AP	1	1	0	0	0	0	0	0
EGFR: AMP	6	12	0	0	3	2	0	1
FGFR1:AMP	76	105	5	2	8	9	9	11
RHOA:AMP	1	0	0	0	0	0	0	0
AKT2:AMP	7	3	0	0	2	2	0	1
FGFR3:AMP	5	5	0	0	0	1	0	1
STK11:AMP	0	2	0	0	0	0	0	0

(continued)

Table 3 (continued)

	Breast cancer subtype							
Amplified Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
TSC2:AMP	3	6	0	0	0	0	0	0
RASA1:AMP	3	0	0	0	1	0	1	0
AKT3:AMP	6	10	1	1	1	0	0	1
FGFR2:AMP	5	16	1	1	2	0	2	2
NRAS:AMP	2	0	0	0	0	1	0	0
AKT1:AMP	2	5	0	0	0	0	0	1
HRAS:AMP	0	0	0	0	0	0	0	1
TSC1:AMP	0	0	0	0	0	0	0	0

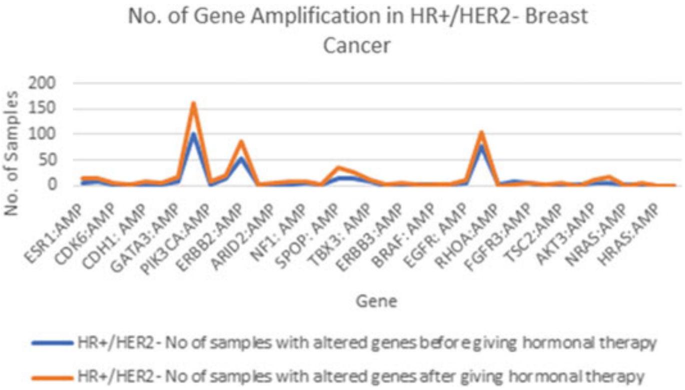


Fig. 7 Gene amplification in HR+/HER2-Breast Cancer

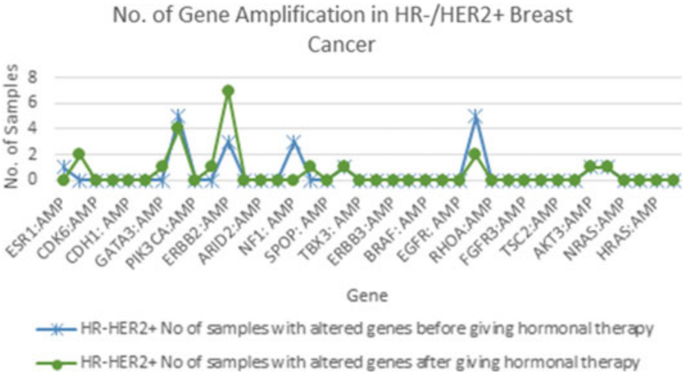


Fig. 8 Gene amplification in HR-/HER2+ Breast Cancer

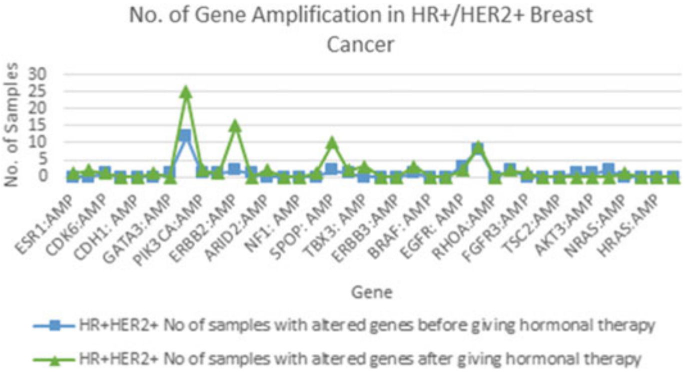


Fig. 9 Gene amplification in HR+/HER2+ Breast Cancer

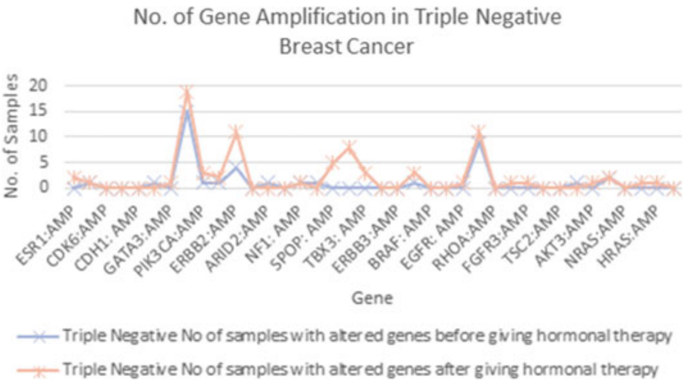


Fig. 10 Gene amplification in triple negative Breast Cancer

metastases breast cancer some genes have shown increased alterations. Also, details of amplified genes, deleted genes and fusion genes are also available. Due to failure of therapy amplification of some of the genes is found. Also, deletion in some of the genes is also found.

ESR1:

This gene encodes an estrogen receptor. The receptor encoded by this gene plays a vital role in breast cancer. The identification of ER-positive breast cancers that are resistant to hormonal therapy can be detected with the help of ESR1 mutations. ESR1 is a protein coding gene. Breast cancer along with estrogen resistance is a disease associated with ESR1 gene [17]. ESR1 gene encodes estrogen receptor protein. Studies show that ESR1 mutations are found in tumors which have shown resistance to hormonal therapy. ESR1 mutation plays a key role, as biomarker in personalized therapy recommendation. Yet, complete association of this gene mutation with treatment failure is not fully discovered and has a huge research scope in terms of personalized therapy recommendation [21].

CDH1:

It is a tumor suppressor gene. This gene encodes a classical cadherin. Mutations in this gene are correlated with gastric, breast, colorectal, thyroid and ovarian cancer. Loss of function of this gene contributes to cancer progression by increasing proliferation, metastases.

Around 10% invasive breast cancers are invasive lobular carcinomas. Invasive lobular carcinoma is found to be strongly associated with female hormones.

CDH1 gene is initially correlated with gastric cancer but lobular type breast cancers are also found to be associated with this gene. The risk of developing cancer in females with CDH1 mutation is nearly 50% [22].

About 3% of the studied samples in women had lobular breast carcinoma, and it was also identified as hereditary breast cancer [23].

TP53:

Table 4 Frequency of gene alteration in deleted genes before and after hormonal therapy

	Breast cancer subtype							
Deleted Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
ESR1: HOMDEL	4	7	3	0	0	0	1	1
CDK4: HOMDEL	7	13	0	2	0	2	1	1
CDK6: HOMDEL	2	4	0	0	1	1	0	0
MTOR: HOMDEL	1	2	0	0	0	0	0	0
CDH1: HOMDEL	3	7	0	0	0	0	0	0
TP53: HOMDEL	3	4	0	0	0	1	1	0
GATA3: HOMDEL	9	18	0	1	1	0	0	1
CCND1: HOMDEL	101	163	5	4	12	25	15	19
PIK3CA: HOMDEL	3	8	0	0	1	2	1	3

(continued)

Table 4 (continued)

Deleted Gene	Breast cancer subtype							
	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
PTEN: HOMDEL	13	19	0	1	1	1	1	2
ERBB2: HOMDEL	52	85	3	7	2	15	4	11
ARID1A: HOMDEL	1	1	0	0	1	0	0	0
ARID2: HOMDEL	2	6	0	0	0	2	1	0
CREBBP: HOMDEL	3	9	0	0	0	0	0	0
NF1: HOMDEL	4	7	3	0	0	0	1	1
FGFR4: HOMDEL	2	2	0	1	0	1	1	0

(continued)

Table 4 (continued)

Deleted Gene	Breast cancer subtype							
	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
SPOP: HOMDEL	13	36	0	0	2	10	0	5
FOXA1: HOMDEL	14	27	1	1	1	2	0	8
TBX3: HOMDEL	7	11	0	0	0	3	0	3
CTCF: HOMDEL	0	1	0	0	0	0	0	0
ERBB3: HOMDEL	1	4	0	0	0	0	0	0
KRAS: HOMDEL	2	3	0	0	1	3	1	3
BRAF: HOMDEL	3	1	0	0	0	0	0	0

(continued)

Table 4 (continued)

	Breast cancer subtype							
Deleted Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
MAP2K1: HOMDEL	1	1	0	0	0	0	0	0
EGFR: HOMDEL	6	12	0	0	3	2	0	1
FGFR1: HOMDEL	76	105	5	2	8	9	9	11
RHOA: HOMDEL	1	0	0	0	0	0	0	0
AKT2: HOMDEL	7	3	0	0	2	2	0	1
FGFR3: HOMDEL	5	5	0	0	0	1	0	1
STK11: HOMDEL	0	2	0	0	0	0	0	0

(continued)

Table 4 (continued)

	Breast cancer subtype							
Deleted Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
TSC2: HOMDEL	3	6	0	0	0	0	0	0
RASA1: HOMDEL	3	0	0	0	1	0	1	0
AKT3: HOMDEL	6	10	1	1	1	0	0	1
FGFR2: HOMDEL	5	16	1	1	2	0	2	2
NRAS: HOMDEL	2	0	0	0	0	1	0	0
AKT1: HOMDEL	2	5	0	0	0	0	0	1
HRAS: HOMDEL	0	0	0	0	0	0	0	1

(continued)

Table 4 (continued)

	Breast cancer subtype							
Deleted Gene	HR+/HER2-		HR-HER2+		HR+HER2+		Triple Negative	
	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy	No of samples with altered genes before giving hormonal therapy	No of samples with altered genes after giving hormonal therapy
TSC1: HOMDEL	0	0	0	0	0	0	0	0

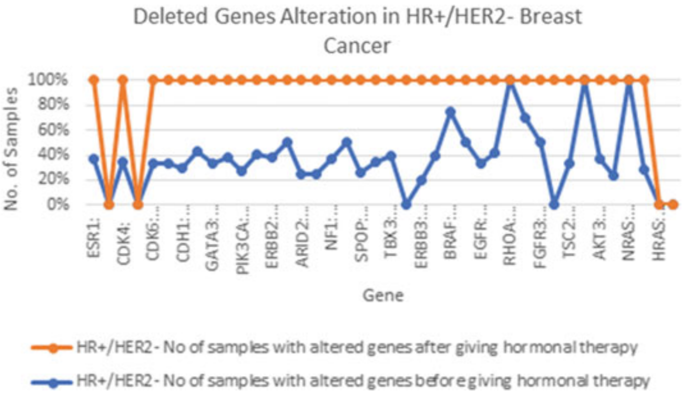


Fig. 11 Gene deletion in HR+/HER2-Breast Cancer

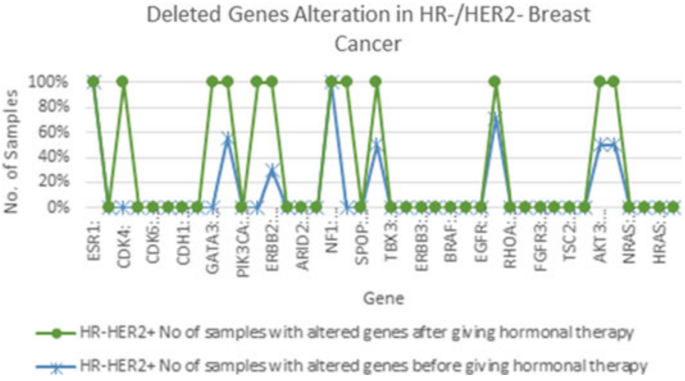


Fig. 12 Gene amplification in HR-/HER2-Breast Cancer

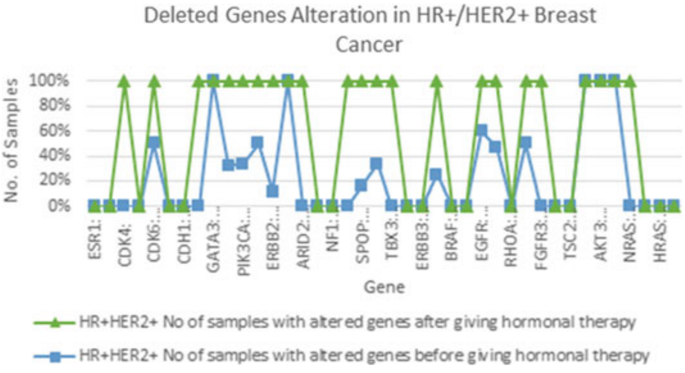


Fig. 13 Gene deletion in HR+/HER2_ Breast Cancer

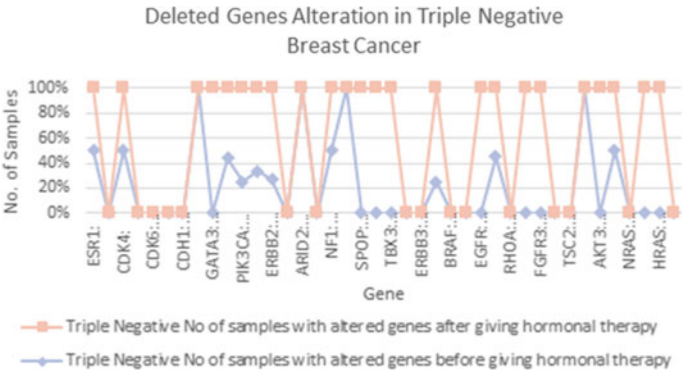


Fig. 14 Gene deletion in triple negative Breast Cancer

This gene encodes a tumor suppressing protein. Mutations in these genes are associated with a variety of human cancers including hereditary cancers. TP53 is a protein coding gene.

In breast cancers, the TP53 gene is mutated in nearly 30% of the patients. Mutation in TP53 can be a good factor of prognosis in some cases while it can be a poor factor of prognosis in other cases.

Variety of therapies are available for breast cancer like chemotherapy, radiation therapy, hormone therapy, etc.

Generally, advanced metastases breast cancers are treated with hormone therapy.

Most cancers associated with TP53 mutations have shown good response to anthracycline chemotherapy, no anthracycline chemotherapy and radio therapy. Studies show that mutations in TP53 genes are not correlated with hormonal therapy and hence shall be avoided as it may resist the treatment [24].

GATA3:

This gene encodes a protein which belongs to the GATA family of transcription factors. This is a protein coding gene. GATA3 is closely correlated with estrogen receptor type breast cancers.

As GATA3 is associated with ER-breast cancer, studies show that ER-negative tumors respond to endocrine therapies. So, this gene can be used as a biomarker in case of ER-negative breast cancer. It can be concluded that this gene has no correlation with endocrine or hormonal therapy, and hence, tumors with GATA3 mutations can be given treatment other than hormone therapy [25].

CCND1:

CCND1 is a protein coding gene. Diseases associated with this gene includes cancer.

CCND1 is amplified in primary breast cancers and also ER-positive breast cancers. Studies show that breast cancers with amplification of CCND1 show poor response to endocrine therapy [26].

PTEN:

This is a tumor suppressor gene and found to be mutated in large numbers of cancers with high frequency. It is mutated in prostate cancer, glioblastoma, endometrial, lung and breast cancer with varying amounts.

Studies show that PTEN acts as a negative regulator in ER-negative breast cancers. Reduced PTEN levels result in resistance to endocrine therapy [27].

ERBB2:

ERBB2 gene is commonly referred to as HER2. It is always amplified and overexpressed in 20–30% breast invasive carcinomas. ERBB2 activating mutations have been shown to have clinical importance in HER2-negative breast cancer.

Studies show that HER2-positive breast cancer may show resistance to endocrine therapy [28].

FGFR1:

FGFR1 is a protein coding gene. It is involved in many types of cancers in the form of amplification, overexpression, etc.

Amplification of FGFR1 is found in nearly 10% of breast cancers.

FGFR1 overexpression promotes resistance to endocrine therapy [29].

FOXA1:

FOXA1 is a protein coding gene. Diseases associated with FOXA1 include ER-positive and ER-negative breast cancer.

The findings show that FOXA1 upregulation results in resistance to endocrine therapy in metastatic breast cancer [30].

PIK3CA:

Approximately, 30% of all patients carry mutations of the PIK3CA gene. Mutations in PIK3CA are associated with resistance to endocrine therapy, HER-2 directed therapy and cytotoxic therapy. Study shows that mutations in PIK3CA can be considered as a biomarker in deciding treatment for HR+HER2_ advanced breast cancers [31].

NF1:

NF1 is a protein coding gene and associated with cancer including many other diseases.

Studies found that NF1 alterations are found in invasive lobular carcinoma. Studies show that loss in NF1 results in endocrine therapy resistance [32].

5 Conclusion

The advanced metastases breast cancer is a heterogeneous type cancer. They show different behavior to different therapies. In order to increase chances of survival, most accurate treatment has to be given. Looking at the complexity of the disease, molecular profiling of tumor samples has to be done in order to understand the disease in terms of genetic variations. Once the genomic landscape of the tumor is understood, clinicians can refer to the research which has discovered biomarkers in cancer type. Knowing these biomarkers, doctors can recommend suitable treatment to the patient.

With the same aim, the proposed study explores the hormone therapy failure dataset of advanced metastases breast cancer. The study focuses on genomic aspects of the tumor samples.

Genes are altered in cancer in the form of mutation, amplification, deletion and fusion. All genomic alterations were studied with the help of present study and literature review.

It is observed that there is an increase in ESR1, CDH1, TP53, GATA3, PIK3CA, PTEN, ERBB2, FOXA1 and FGFR1 mutations after hormonal therapy. CCND1 gene is seen to be amplified after hormonal therapy. NF1 gene loss is seen after hormonal therapy.

The studied genes can be used as biomarkers in therapy recommendation. It can be concluded that if such changes in genes are present, hormonal therapy is not suitable for the patients.

References

1. U.S. Breast Cancer Statistics. www.breastcancer.org. Accessed 4 Sept 2021
2. Sharma GN et al (2010) Various types and management of breast cancer: an overview. J Adv Pharmaceutical Technol Res 1,2: 109–126
3. Alkabban FM, Ferguson T. Breast Cancer. [Updated 2021 Aug 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK482286/
4. Harbeck N, Penault-Llorca F, Cortes J et al (2019) Breast cancer. Nat Rev Dis Primers 5**:** 66. https://doi.org/10.1038/s41572-019-0111-2
5. Bonotto M et al (2015) Treatment of metastatic breast cancer in a real-world scenario: is progression-free survival with first line predictive of benefit from second and later lines? The Oncologist 20(7:719–724. https://doi.org/10.1634/theoncologist.2015-0002
6. Roché H, Vahdat LT (2011 May) Treatment of metastatic breast cancer: second line and beyond. Ann Oncol 22(5):1000–1010. https://doi.org/10.1093/annonc/mdq429 Epub 2010 Oct 21 PMID: 20966181
7. Sun Y-S et al (2017) Risk factors and preventions of breast cancer. Int J Biol Sci 13(11):1387–1397. https://doi.org/10.7150/ijbs.21635
8. Krassowski M, Das V, Sahu SK, Misra BB (2020) State of the field in multi-omics research: from computational needs to data mining and sharing. Front Genet 10(11):610798. https://doi.org/10.3389/fgene.2020.610798.PMID:33362867;PMCID:PMC7758509
9. Gambardella V, Tarazona N, Cejalvo JM, Lombardi P, Huerta M, Roselló S, Fleitas T, Roda D, Cervantes A (2020) Personalized medicine: recent progress in cancer therapy. Cancers (Basel) 12(4):1009. https://doi.org/10.3390/cancers12041009.PMID:32325878;PMCID:PMC7226371
10. Dai X et al (2015) Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res 5(10):2929–2943
11. Bertos NR, Park M (2011) Breast cancer—one term, many entities? J Clinical Investigation 121(10:3789–3796. https://doi.org/10.1172/JCI57100
12. My Cancer Genome. https://www.mycancergenome.org/content/disease/breast-invasive-ductal-carcinoma/. Accessed 5 Sept 2021
13. Ellis MJ, Perou CM (2013) The genomic landscape of breast cancer as a therapeutic roadmap. Cancer Discovery 3(1):27–34. https://doi.org/10.1158/2159-8290.CD-12-0462

14. Schwab RB, Koehler M, Ali SM, Murray BW (2016) Genomic profiling and treatment of HER2+, ER+, PgR+ “triple positive” breast cancer: a case report and literature review. Cancer Treatment Res Commun 9:27–31. ISSN 2468-2942. https://doi.org/10.1016/j.ctarc.2016.06.008
15. Ramón y Cajal S, Sesé M, Capdevila C et al (2020) Clinical implications of intratumor heterogeneity: challenges and opportunities. J Mol Med 98**:**161–177. https://doi.org/10.1007/s00109-020-01874-2
16. Natrajan R et al (2016) Microenvironmental heterogeneity parallels breast cancer progression: a histology-genomic integration analysis. PLoS Med 13(2):e1001961. https://doi.org/10.1371/journal.pmed.1001961
17. Razavi P et al (2018) The genomic landscape of endocrine-resistant advanced breast cancers. Cancer Cell 34(3):427–438.e6. https://doi.org/10.1016/j.ccell.2018.08.008
18. Custers R et al (2019) Genetic alterations that do or do not occur naturally; Consequences for genome edited organisms in the context of regulatory oversight. Front Bioeng Biotechnol 6:213. https://doi.org/10.3389/fbioe.2018.00213
19. Kadota M, Sato M, Duncan B, Ooshima A, Yang HH, Diaz-Meyer N, Gere S, Kageyama S-I, Fukuoka J, Nagata T, Tsukada K, Dunn BK, Wakefield LM, Lee MP (2009) Identification of novel gene amplifications in breast cancer and coexistence of gene amplification with an activating mutation of *PIK3CA*. Cancer Res 69(18):7357–7365. https://doi.org/10.1158/0008-5472.CAN-09-0064
20. Dischinger PS, Tovar EA, Essenburg CJ et al (2018) *NF1* deficiency correlates with estrogen receptor signaling and diminished survival in breast cancer. npj Breast Cancer 4:29. https://doi.org/10.1038/s41523-018-0080-8
21. Reinert T, Gonçalves R, Bines J (2018) Implications of ESR1 mutations in hormone receptor-positive breast cancer. Curr Treat Options Oncol 19(5):24. https://doi.org/10.1007/s11864-018-0542-0 PMID: 29666928
22. Dossus L, Benusiglio PR (2015) Lobular breast cancer: incidence and genetic and non-genetic risk factors. Breast Cancer Res: BCR 17:37. https://doi.org/10.1186/s13058-015-0546-7
23. Corso G et al (2018) Prognosis and outcome in CDH1-mutant lobular breast cancer. Eur J Cancer Prevention: Official J Eur Cancer Prevention Organisation (ECP) 27(3):237–238. https://doi.org/10.1097/CEJ.0000000000000405
24. Varna M et al (2011) TP53 status and response to treatment in breast cancers. J Biomed Biotechnol 2011: 284584. https://doi.org/10.1155/2011/284584
25. Albergaria A, Paredes J, Sousa B et al (2009) Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. Breast Cancer Res 11:R40. https://doi.org/10.1186/bcr2327
26. Li Z et al (2016) Evaluation of CCND1 amplification and CyclinD1 expression: diffuse and strong staining of CyclinD1 could have same predictive roles as CCND1 amplification in ER positive breast cancers. Am J Translational Res 8(1):142–153
27. Fu X, Creighton CJ, Biswal NC, Kumar V, Shea M, Herrera S, Contreras A, Gutierrez C, Wang T, Nanda S, Giuliano M, Morrison G, Nardone A, Karlin KL, Westbrook TF, Heiser LM, Anur P, Spellman P, Guichard SM, Smith PD, Schiff R et al (2014) Overcoming endocrine resistance due to reduced PTEN levels in estrogen receptor-positive breast cancer by co-targeting mammalian target of rapamycin, protein kinase B, or mitogen-activated protein kinase kinase. Breast Cancer Res: BCR 16(5):430. https://doi.org/10.1186/s13058-014-0430-x
28. Wang J, Xu B (2019) Targeted therapeutic options and future perspectives for HER2-positive breast cancer. Sig Transduct Target Ther 4:34. https://doi.org/10.1038/s41392-019-0069-2
29. Turner N, Pearson A, Sharpe R, Lambros M, Geyer F, Lopez-Garcia MA, Natrajan R, Marchio C, Iorns E, Mackay A, Gillett C, Grigoriadis A, Tutt A, Reis-Filho JS, Ashworth A (2010) FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. Cancer Res 70(5):2085–2094. https://doi.org/10.1158/0008-5472.CAN-09-3746. Epub 2010 Feb 23. PMID: 20179196; PMCID: PMC2832818
30. Fu X, Pereira R, De Angelis C, Veeraraghavan J, Nanda S, Qin L, Cataldo ML, Sethunath V, Mehravaran S, Gutierrez C, Chamness GC, Feng Q, O’Malley BW, Selenica P, Weigelt B, Reis-Filho JS, Cohen O, Wagle N, Nardone A, Jeselsohn R, Brown M, Rimawi MF, Osborne CK,

Schiff R (2019) FOXA1 upregulation promotes enhancer and transcriptional reprogramming in endocrine-resistant breast cancer. Proc Natl Acad Sci U S A 116(52):26823–26834. https://doi.org/10.1073/pnas.1911584116. Epub ahead of print. PMID: 31826955; PMCID: PMC6936436

31. Schwartzberg LS, Vidal GA (2020) Targeting PIK3CA alterations in hormone receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: new therapeutic approaches and practical considerations. Clin Breast Cancer 20(4):e439–e449. https://doi.org/10.1016/j.clbc.2020.02.002 Epub 2020 Feb 20 PMID: 32278641
32. Sokol ES et al (2019) Loss of function of NF1 is a mechanism of acquired resistance to endocrine therapy in lobular breast cancer. Ann Oncology : Official J Eur Soc Med Oncol 30(1):115–123. https://doi.org/10.1093/annonc/mdy497

Chapter 4
TrueReview—A Review Scraper and Analyzer for Customer Product Reviews for E-commerce Sites using Text Mining

Sangeeta Kumari, Shubham Mathesul, Ayush Rambhad, and Abrar Momin

1 Introduction

The exponential expansion of ICT has reduced the dependence on human experts and books and has made it simpler for people to acquire what they desire. One of the key factors attributing to the advancement of product review analysis is sentiment analysis. Sentiment analysis is the calculation of people's sensations, their opinions, emotions, and sentiments in relation to a entity. In today's world, online shopping is the go-to choice for most customers. Usually, manufacturers offer their Websites with information about the product, however, they focus more on the key features and quality of the product from their perspective rather than the user's perspective of the product. More information about these consumer items is also available on many popular Internet sites. Online consumers are growing rapidly every day, and there is a huge variety of product reviews available for many product categories. Although reviews are made easily accessible by many Websites, a potential buyer is unable to read all reviews for a certain type of product as hundreds or even thousands of reviews are available for each product [1]. Product reviews are an important element of the branding and marketing of any company be it a physical store or an online company. Consumer's opinion of a product makes or breaks the online perception of the product as well as the company. Today's generation heavily relies on these

S. Kumari · S. Mathesul (✉) · A. Rambhad · A. Momin
Department of Computer Engineering, Vishwakarma Institute of Technology, Pune 411037, India
e-mail: shubham.mathesul19@vit.edu

S. Kumari
e-mail: sangeeta.kumari@vit.edu

A. Rambhad
e-mail: ayush.rambhad19@vit.edu

A. Momin
e-mail: abrar.momin19@vit.edu

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_4

reviews to influence their purchase decision. Each consumer wants to know the exact in-depth details about the key features of a product. They often try to compare two or more product sharing the same features to pinpoint the product, they finally choose to purchase, and reviews play a key role in providing the basis for comparison based on the previous usage by another consumer. Given the fact that there are hundreds and thousands of products with consumers and a huge data collection of their reviews, it is challenging for customers and manufacturers to monitor and determine consumer sentiments each time. We thus split product reviews according to the key features of the items. Each feature is classified as bad or good and offers an indicator to the manufacturer as well as the consumer. Professional analysts study the categories and determine which characteristics influence more in the sales of these items and which add to their decline. We provide a final score in terms of percentage indicating to either how positive or how negative the reviews were to provide a detailed analysis rather than focusing on the ratings of the product. Finally, the many features of the product that increase its popularity or demographic are shown and analyzed visually according to each individual feature with the same performed for the declining area of the product. Using the same analysis performed till now, we extract the top 20 words used by the consumers in their reviews, this takes user convenience to a whole new level as it eliminates the need of the user to read each every to find the principal indication of it. This helps highlight the popular and unpopular features of the product with the adjectives used to describe their sentiment of the product.

2 Literature Survey

See Table 1.

3 Methodologies

In this paper, we present a method for the review scraper which extracts user reviews from consumer goods. Review sites contain online product reviews which include unstructured and non-refined texts, which reflect opinions from several online users. The system is divided into three major phases. The system structure is shown in Fig. 1.

The first stage is the crawling phase in which product reviews from numerous consumer product reviews sites are collected. The second phase is the analysis in which scraped reviews are parsed, preprocessed, and evaluated for relevant information to be extracted or obtained from it. The third phase involves classifying and evaluating review datasets with several accessible classifiers.

Table 1 Literature review

Paper title	Authors	Objective	Methodology	Accuracy	Outcome
E-commerce product review sentiment classification based on a Naïve Bayes continuous learning framework [2]	Xu et al. [2]	To improve the knowledge-based distribution on three types of hypothesis to better adapt to various domains [2]	Naïve Bayes	80%	To demonstrate that, in addition to the benefit of computational efficiency, the old and new methods have been properly traded
Product review analysis for genuine rating [3]	Sagar et al. [3]	To design a sophisticated system rating sentiment analysis which uncovers hidden feelings in the commentary of the user [3]	VaderSharp	NA	Users reviews and hidden thoughts were analyzed, and the text was classified as positive and negative by utilizing VaderSharp
Review analyzer: analyzing consumer product Reviews from review collections [4]	Arun Manicka Raja et al. [4]	Ability to determine the polarity of the review and to perform review classification [4]	KNN, Naïve Bayes	Test accuracy of 0.795	Results include classification of reviews and a comparison of the system with existing classifiers
Sentiment analysis on large scale Amazon product reviews [5]	Haque et al. [5]	To use ML methods to generate satisfactory polarity results from Amazon user reviews [5]	SVM	Tenfold accuracy of 93.57%	Supervised learning model that outputs polarity and uses a combination of 2 feature extraction methods to improve accuracy
Research on product review analysis and spam review detection [6]	Chauhan et al. [6]	Attempt to detect spam and false reviews and filter reviews containing profane or swear words by adding a study of sentiment [6]	VaderSharp	57.2%	Weight according to its polarity provided to the word and the sentiment score after NLP processing

(continued)

Table 1 (continued)

Paper title	Authors	Objective	Methodology	Accuracy	Outcome
Modeling and prediction of online product review helpfulness: a survey [7]	Diaz and Ng [7]	To study the elements determining helpfulness of the review and tries to anticipate it properly [7]	NA	NA	An summary of the most important studies on the prediction and comprehension of product review helpfulness

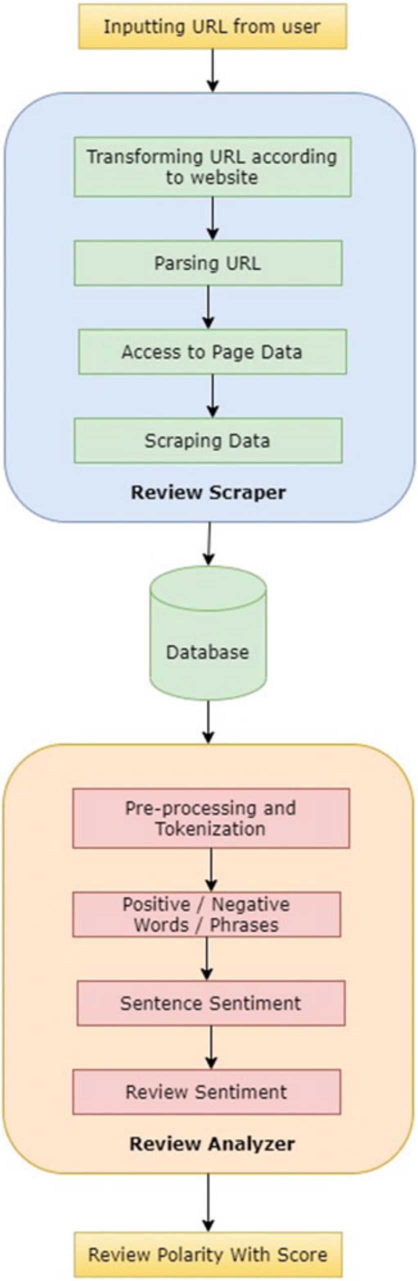
3.1 *Crawling the Web*

Many consumers take purchase decisions by taking into account other user's opinions and concerns. For instance, most consumers will be purchasing the goods which are ranked most popular or highly rated. Therefore, user opinions must be crawled and processed to enable other consumers to make use of them and influence online goods decision-making procedures. So, to perform Web crawling, we designed our own automated Web crawler. It is a search engine for real-time user reviews that helps to collect a single URL from the user and store the gathered URLs in a.csv file. This file is then passed to the crawler to visit each of these URL and extract all the information present on the Webpage. We crawled the Webpages for the most common attributes present on the page—product name, manufacturer, user rating, all the user reviews along with the date-time information. We use an XML schema which specifies the structure and the data, components, or attributes of an XML document that can be included.

3.2 *Preprocessing Crawled Data*

Product reviews might contain profanity and swear language. More advanced noise cleaning methods in raw text are needed in order to correctly classify the reviews and prevent the augmentation of reviews to one sentiment due to the use of expletives. We constructed a collection of words with about 100+ expression words, plus other terms that are non-relevant. It involves removal of white spaces, handling of special characters like (^, #, etc.), handling of punctuations and merging the date-time formats into a single standard format.

Fig. 1 TrueReview architecture



3.3 Review Classification

Our scraped Amazon reviews dataset ranks various goods on a scale of 1–5. Then, we build our own lexicon with keywords/adjectives which can explain user's sentiments in a review in order to get weightage of the review and a sentiment score. The sentiment score reflects the polarity of reviews. It helps the system starts with a judgment for classification in either the positive/negative state. By including the sentiment weightage of each word in the review, we generated the sentiment score. The score is in the -1 to 1 scale. Let us take into account an example review of "excellent gaming performance but a poor battery life." Here, the words "excellent" and "poor" both express a certain feeling. In order to compute the overall sentiment score, we match the weights of sentiment defining terms to the words in our dictionary. The equation used to calculate the final polarity scores is a result of "lambda x:compound_score(x)."

While the overall sentiment score is being calculated, we need to take into consideration the negative prefixes, words such as adjectives and verbs can communicate the opposite feeling in the reviews. Take into account the review of "average gaming performance but exceptional pricing and marketing done by company." Here, the word "exceptional" indicates a positive sentiment, however, the context of the usage of this word displays a negative sentiment. Hence, processing of negative prefixes in adjectives and verbs performs a crucial role while calculating the overall sentiment score. Once the sentiment score is calculated, in order to assign classified labels for the input review, the Naïve Bayes classification method is utilized. The classifier is trained using the review dataset, and the same review datasets are used for testing. In the training review dataset, Naïve Bayes calculates the preceding probability of each label and outputs the final classified label for the input review (Fig. 2).

On the homepage of the application, an input textbox takes the URL of the product he wishes to find the review. Currently, the Website support is limited to Amazon. Then, the system is configured to start scraping the reviews present on the review page from the URLs. It scrapes all the reviews in real-time and sends it to the classifier to find the discriminative review of the product. For the scraping process, we make use of an automation framework called BeautifulSoup. Then, the input is preprocessed, filtered, and tokenized so that we can perform natural language processing on it to help get a clear idea of the review polarity. Each review will be given a rating between 1 and 5 along with a score which helps determine the final rating score which decides

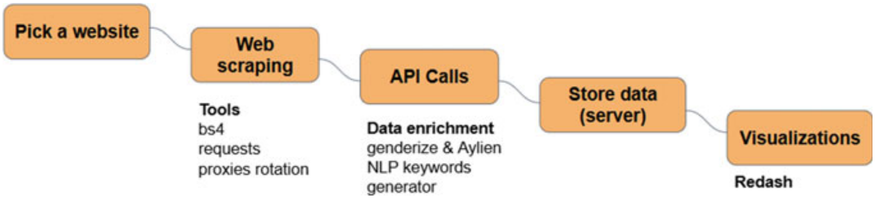


Fig. 2 TrueReview processing flow

the polarity of the review. All of this data are stored locally on the server which can be downloaded by the administrators for verification. Finally, the data which are stored are sent to the analysis system to generate various visualization out of it, all of which are displayed in the next section.

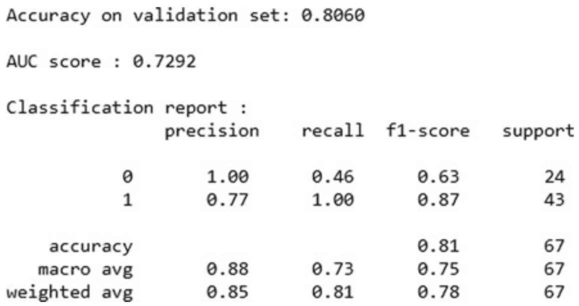
4 Results

A variety of consumer product reviews have been provided by many product Websites. To train our classifier on these reviews, we created the dataset of 5065+ reviews from all the different product categories of the Amazon Website. Experimental results show that the review scraper takes minimal time to scrape all the reviews in real-time and sends it to the classifier to find the visually discriminative review of the product. The tests are provided by explaining the procedure for processing the examination datasets of the review scraper system and calculating the classified findings of the review analysis system. The Naïve Bayes algorithm for classification groups reviews depending on the direction of the feeling. It analyzes all the keywords with their contexts, detects the sentiment-oriented terms in the classification contents, and then outputs the classification label according to polarity. TrueReview sentiment classification achieved an accuracy of **80.60%** on the test dataset (Figs. 3 and 4).

Once the user browses to the homepage of the application, an input textbox is made available to the user for inputting the URL of the product he wishes to find the review of from TrueReview. The product URL used for this example: https://www.amazon.in/Apple-iPhone-Pro-Max-512GB/dp/B07XVLCSHP/ (Fig. 5).

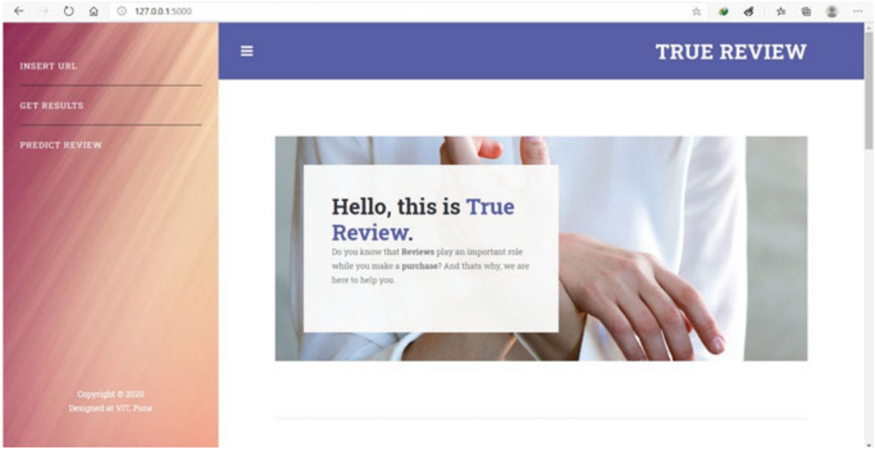
The user is then provided with the word cloud of the top keywords used in the review which helps user briefly understand the key features and popular opinions on the product.

Figure 6 depicts plot graph feature of TrueReview that helps user to understand the major polarity of the sentiments present in the reviews. The plot consists of 3 polls—users with negative sentiment (blue), users with positive sentiment (orange), and users with neutral sentiments (green).



```
Accuracy on validation set: 0.8060

AUC score : 0.7292

Classification report :
              precision    recall  f1-score   support

           0       1.00      0.46      0.63        24
           1       0.77      1.00      0.87        43

    accuracy                           0.81        67
   macro avg       0.88      0.73      0.75        67
weighted avg       0.85      0.81      0.78        67
```

Fig. 3 TrueReview evaluation results

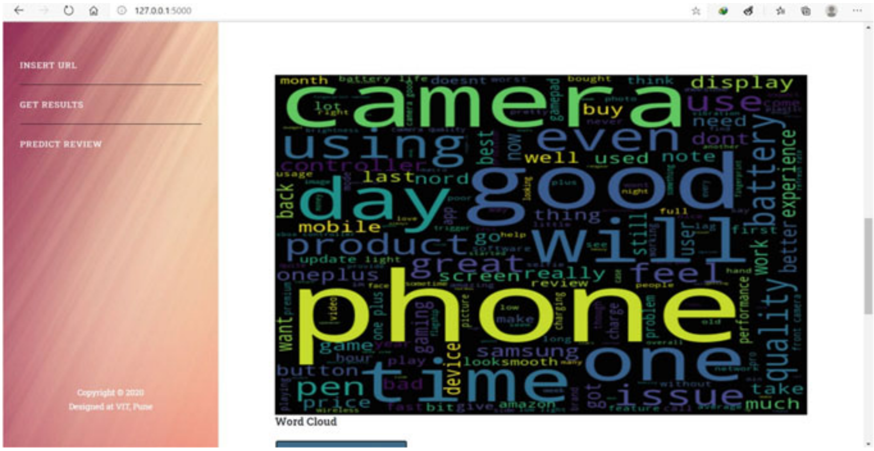


Hello, this is True Review.

Do you know that **Reviews** play an important role while you make a **purchase**? And thats why, we are here to help you.

Copyright © 2020
Designed at VIT, Pune

Fig. 4 TrueReview homepage



Copyright © 2020
Designed at VIT, Pune

Fig. 5 TrueReview word cloud

Unigrams which are the collection of top ranking words are made available to the user to understand the key features the users like (blue) and disliked (red) about the product as shown in Fig. 7.

One of the central principal features of TrueReview is the review sentiment predictor. This tool can be used by users to get started with writing their reviews and to understand how their review contributes to the change in the overall sentiment of the product. The feature has the ability to accurately understand the context of the adjectives used in the review. This tool performs the classification of the user input review as shown in Fig. 8 and helps users understand how they can change their review to make it influence positively/negatively.

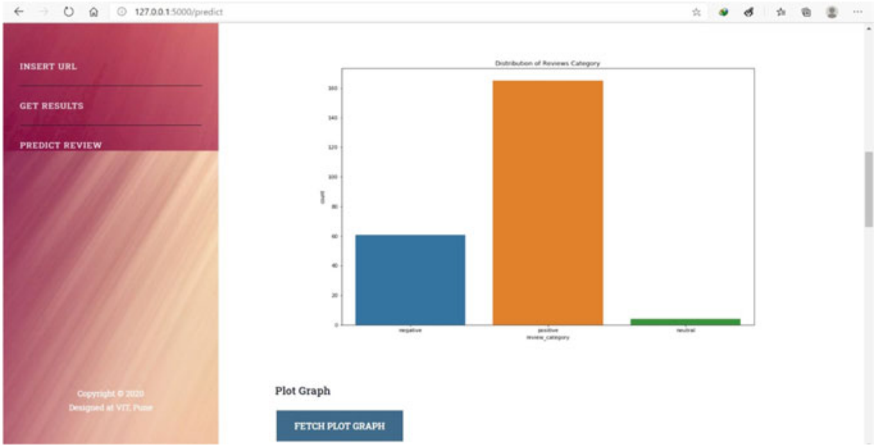


Fig. 6 TrueReview review polarity graph

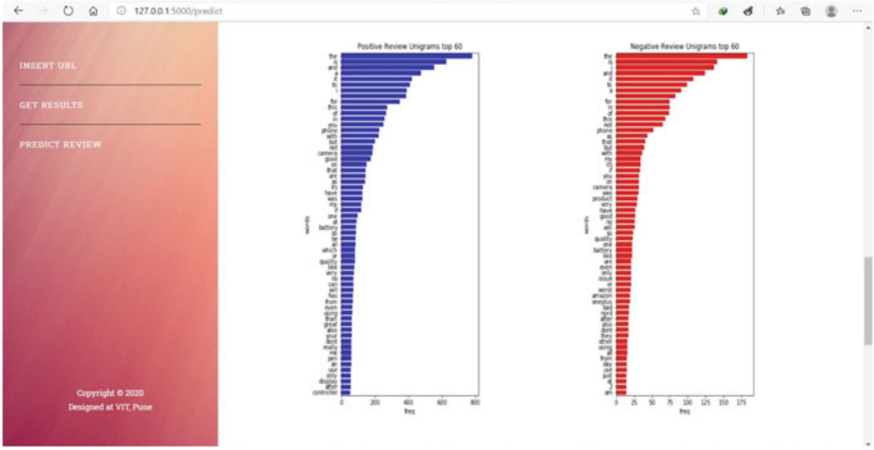
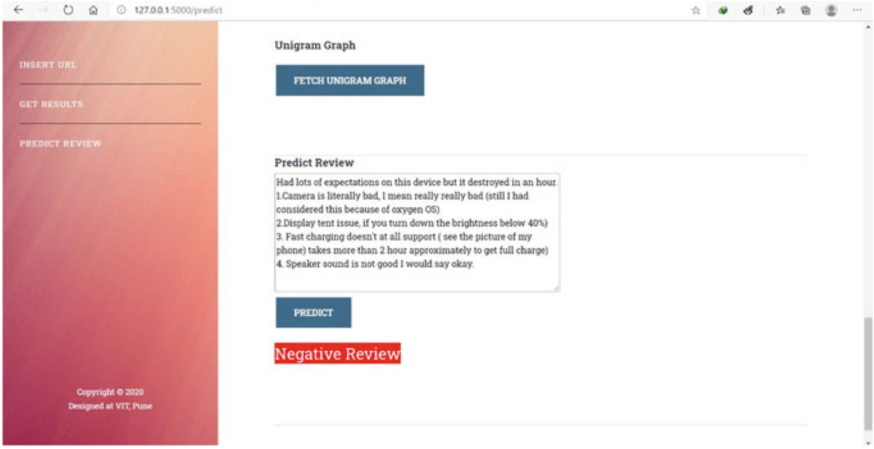


Fig. 7 TrueReview unigram

5 Conclusion

In this research, we introduced TrueReview, a review scraper and analyzer to help online shopping consumer ease their way into understanding the TrueReview of the product. The research has shown that many users often find it difficult to read all the reviews present in the product page and often get influenced by the top reviews present on the page. To automate the processing of analysis of the real value of the product, we designed this application so that the user finds it convenient to just provide the URL of the product they are interested in and finds a visually discriminative and methodical analysis of the product in real-time. Though the algorithm achieved an



Had lots of expectations on this device but it destroyed in an hour.
1.Camera is literally bad, I mean really really bad (still I had considered this because of oxygen OS)
2.Display tent issue, if you turn down the brightness below 40%)
3. Fast charging doesn't at all support (see the picture of my phone) takes more than 2 hour approximately to get full charge)
4. Speaker sound is not good I would say okay.

Negative Review

Fig. 8 TrueReview review sentiment predictor

accuracy of 80.60% on the test dataset, the system can be further refined by using advanced preprocessing techniques for processing the crawled data to find improved sentiment scores and the overall accuracy of the system can be improved by addition of the collection of unidentified keywords and updating the weights associated with each of the keywords.

References

1. Sultana N, Kumar P, Patra M, Chandra S, Alam S (2019) Sentiment analysis for product review. Int J Soft Comput 09:7. https://doi.org/10.21917/ijsc.2019.0266
2. Xu F, Pan Z, Xia R (2020) E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. Inf Process Manage 57(5):102221. ISSN 0306-4573. https://doi.org/10.1016/j.ipm.2020.102221
3. Sagar P, Anju P, Vinit S, Nikhilesh J (2020) Product review analysis for genuine rating. Int Res J Eng Technol (IRJET) e-ISSN: 2395–0056
4. Arun Manicka Raja M, Godfrey Winster S, Swamynathan S (2012)Review analyzer: analyzing consumer product reviews from review collections. In: International conference on recent advances in computing and software systems, pp 287–292. https://doi.org/10.1109/RACSS.2012.6212682
5. Haque TU, Saber NN, Shah FM (2018) Sentiment analysis on large scale Amazon product reviews. In: IEEE international conference on innovative research and development (ICIRD), pp 1–6.https://doi.org/10.1109/ICIRD.2018.8376299
6. Chauhan SK, Goel A, Goel P, Chauhan A, Gurve MK (2017) Research on product review analysis and spam review detection. In: 4th International conference on signal processing and integrated networks (SPIN), pp 390–393. https://doi.org/10.1109/SPIN.2017.8049980
7. Diaz G, Ng V (2018) Modeling and prediction of online product review helpfulness: a survey. 698–708. https://doi.org/10.18653/v1/P18-1065
8. Krotov V, Johnson L, Silva L (2020) Tutorial: legality and ethics of web scraping. Commun Assoc Inf Syst 47. https://doi.org/10.17705/1CAIS.04724

9. Dishi J, Bitra V, Saravanakumar K (2019) Sentiment analysis of product reviews a survey. Int J Sci Technol Res 8(12). ISSN 2277–8616
10. Suganya E, Vijayarani S (2020) Sentiment analysis for scraping of product reviews from multiple web pages using machine learning algorithms.https://doi.org/10.1007/978-3-030-16660-1_66
11. Zhao W et al (2018) Weakly-supervised deep embedding for product review sentiment analysis. IEEE Trans Knowl Data Eng 30(01):185–197. https://doi.org/10.1109/TKDE.2017.2756658
12. Sun X, Sun C, Quan C, Ren F, Tian F, Wang K (2017) Fine-grained emotion analysis based on mixed model for product review. Int J Netw Distrib Comput 5:1. https://doi.org/10.2991/ijndc.2017.5.1.1
13. Krutika W, Pranali R, Rushabh B, Nadim B, Bhuvneshwar K (2018) Sentiment analysis of product review. Int J Innovations Eng Sci 3(5). e-ISSN: 2456-3463
14. Jie-Jun L, Han Y, Hao T (2017) Feature mining and sentiment orientation analysis on product review. Manage Inf Optoelectron Eng 79–84
15. Fang X, Zhan J (2015) Sentiment analysis using product review data. J Big Data 2:5. https://doi.org/10.1186/s40537-015-0015-2
16. Prabakaran A, Chen M (2019) Product review credibility analysis. In: International conference on computing, networking and communications (ICNC), pp 11–15. https://doi.org/10.1109/ICCNC.2019.8685490

Chapter 5
Survey of Sentence Scoring Techniques for Extractive Text Summarization

Anushka A. Deshpande and Vinayak G. Kottawar

1 Introduction

Text Summarization is formally defined as technique of collecting important information from an elaborated original text and present it in the form of a summary [1]. Recently, the need for summarization has grown in many domains, and its applications include but are not limited to news articles summarization, email summarization, research papers summarization, medical history summarization and in website summarization to gain information regarding the relevance of that web page to the search made. The primary aim of automatic text summarization is to provide a short summary of a relatively large source text.

Text summarization techniques are broadly classified into two categories: Abstractive and Extractive Summarization [1]. Abstractive Summarization is that in which the source text is read and understood by using linguistic methods. These methods require a deeper understanding of the text but also have the ability to generate new sentences which improve the focus and reduce the redundancy of the text. On the other hand, Extractive Summarization employs various methods to extract sentences and phrases from the source text and groups them together to form a comprehensive summary [1]. This is done without altering any part of the extracted text.

2 Literature Survey

This section describes the related research work in the field of Text Summarization. Text Summarization is broadly classified into two categories, Abstractive and Extractive Summarization. Abstractive Summarization consists of the process of linguistic

A. A. Deshpande (✉) · V. G. Kottawar
D. Y. Patil College of Engineering, Pune, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_5

understanding of the text followed by the generation of a summary. This does not use sentences part of the input text, but, creates new, meaningful sentences for the summary. In Extractive Summarization, sentences from the input text are used as is to form the summary of the text. To perform this operation, each sentence is scored based on certain parameters.

In Madhuri and Ganesh Kumar [2], the author proposes a method for scoring sentences based on the frequency of words. The input file is tokenized into words and stopwords are removed from the corpus. Words not removed are known as keywords and play an important role in summarization. A dictionary of keywords and their frequencies is formed by iterating through the corpus. Weighted frequency is calculated for each sentence by diving the frequency in the sentence by the total frequency in the text. Score of each sentence is the sum of weighted frequencies of the words in that sentence.

In Ramos [3], the concept of Term Frequency—Inverse Document Frequency is proposed for finding word relevance for document queries. TF-IDF method is used for finding the topics of a document. Using these, document retrieval is performed for different queries.

In Shiva Kumar et al. [4], the authors have used the Gensim algorithm to summarize the text. The Gensim algorithm uses multiple techniques such as Word2Vec, FastText, Latent Semantic Indexing, etc. to discover the semantic structure and relationship of the sentences in the document. Based on the results, the sentences are scored and a summary is generated.

3 Extractive Text Summarization Techniques

There are various approaches to perform extractive text summarization. In this subsection, a brief summary of each of these methods is described.

3.1 Statistical-Based Methods

These methods select a sentence based on some assigned weights that determine the relevance for a sentence. The "most important" sentence is the one which has the highest weight assigned [5]. In most of these methods, a value is assigned to each sentence based on the words in that sentence. Multiple algorithms such as word frequency are utilized to assign a value to each sentence.

3.2 Concept-Based Methods

In these methods, concepts are extracted from external knowledge bases. Using these extractions, a vector or graph based model is created to demonstrate a relationship between the concept and sentence [6]. Then, a ranking algorithm is used to extract the most relevant sentences.

3.3 Topic-Based Methods

These methods are dependent on identifying the topic of a document [7]. Methods such as TF-IDF are used to identify the topic of a document. After identifying the topic, each sentence is scored based on its relevance and importance to the topic. Highest scored sentences are then used to form the summary.

3.4 Clustering-Based Methods

These methods are useful in multi-document summarization. All the sentences of each document are placed in a vector space, and clustering algorithms are used to determine similar sentences [8]. The centroid of each cluster, which is a sentence, is used to form the summary of these documents. These methods consider both redundancy and relevance to form the summary.

3.5 Semantic-Based Methods

These methods intend to consider the meaning of the words or sentences to form a summary. LSA is an example of an unsupervised semantic-based method which observes the co-occurrence of words to determine text semantics [9].

3.6 Machine-Learning Based Methods

These transform the problem of text summarization into a supervised learning problem. Models are created to classify each sentence are “summary” or “non-summary” with the help of a training dataset. Neural networks are used to train the model which generates an output score for each sentence [6].

4 Sentence Scoring Algorithms

Extractive text summarization techniques are heavily dependent on sentence scoring algorithms to determine whether a sentence is part of the summary. These algorithms are discussed in this section.

4.1 Word Frequency Algorithm

The word frequency algorithm is the simplest of all algorithms for sentence scoring [2]. A word frequency table is created for the text. Using this table, each sentence is scored. For example,

Consider a sentence S_i with words $W[n]$, then the score of the sentence is (1):

$$\text{Score}[S_i] = \sum \frac{(F[W[i]])}{n} \tag{1}$$

where,

$W[i]$ ith word in the sentence
$F[x]$ frequency of word ‘x’
n number of words in the sentence.

This method gives sufficiently accurate results. But, since this method is heavily dependent on the vocabulary of the document, there is a high chance of irrelevant sentences becoming part of the summary. Another drawback is that semantically similar words have independent effect on the sentence. This implies that words such as “run” and “sprint” are counted separately.

4.2 TF-IDF Algorithm

The TF-IDF approach is best used when there are multiple documents to be summarized into a single summary. TF-IDF considers two very important values as follows:

Term Frequency (TF) of a word is defined as [10],

$$\text{TF} = \frac{\text{number of times } t \text{ occurs in a document}}{\text{Total number of terms in the archive}} \tag{2}$$

The Inverse Document Frequency (IDF) is defined as [10],

$$\text{IDF} = \ln\left(\frac{t}{n}\right) \tag{3}$$

Essentially, TF identifies how incessant a word is, while IDF identifies how a unique a word is. TF-IDF is the score obtained on multiplying these two values [10].

$$\mathrm{TFIDF} = \mathrm{TF} * \mathrm{IDF} \tag{4}$$

where,

t term or word in the document
n number of documents in the archive with the term ‘t’ in them.

TF-IDF is an effective method to generate quick extractive summaries. One of the major downsides of this method is that it can be extremely time consuming in case of large archives that need to be summarized. Also, similar to the word frequency algorithm, this does not take into consideration, the semantic meaning of the word. This can result in imprecise summaries.

4.3 TextRank Algorithm

The TextRank Algorithm is based on the PageRank algorithm used to calculate the weight for web pages [3]. While using the algorithm for TextRank, each web page which was a vertex, now represents each sentence. In this manner, we can find the similarity of two sentences. The formula for calculating the similarity is [11],

$$S(V_i) = (1-d) + d * \sum_{j \varepsilon \mathrm{In}(v_i)} \frac{1}{\left|\mathrm{Out}\left(V_j\right)\right|} S\left(V_j\right) \tag{5}$$

where,

$\mathrm{S(V_i)}$ the weight of the sentence ‘i’
d damping factor, in case of no outgoing links. The default value for this is 0.85
$\mathrm{In}(V_i)$ set of Inbound links to sentence ‘i’
$\mathrm{Out}(V_j)$ set of Outbound links to sentence ‘j’
$|\mathrm{Out}(V_j)|$ number of Outbound links to sentence ‘j’.

The TextRank algorithm can be closely compared with the TF-IDF approach since both algorithms use scoring methods to determine relevance of a word or sentence [3]. The main difference between the two is that TextRank uses the context information of the words to assign the weights. Thus, it assigns comparatively lower scores to words that co-occur only with stop words. On the other hand, TF-IDF simply uses the single word frequency to assign the weight to each word. Thus, it can be said that TextRank weighs the words or sentences with more “strictness” as compared to TF-IDF.

4.4 KL Sum Algorithm

The KL Sum (Kullback-Lieber Sum) algorithm is a sentence scoring algorithm in which the length of the summary is fixed to L words. This method greedily adds sentences to a summary as long as it decreases the KL Divergence [11].

The KL Divergence measures how a probability distribution is different from another. If the divergence is less, it implies that the documents are similar to each other in terms of understanding and meaning conveyed. The KL Divergence is always non-negative and calculated as follows [11]:

$$D_{\mathrm{KL}}(P||Q) \geq 0 \tag{6}$$

$$D_{\mathrm{KL}}(P||Q) = \int_{x_a}^{x_b} P(x) \log \frac{P(x)}{Q(x)} \mathrm{d}x \tag{7}$$

where,

P document set unigram
Q summary distribution
x sentence.

Since this sentence compares the entire document to each sentence to find the KL Divergence, semantic meaning is also taken into account. Through the divergence, relevance of each sentence to the summary is calculated and maintained. Thus, this can give the most accurate summary of all the mentioned algorithms.

5 Text Pre-processing for Each Algorithm

5.1 Word Frequency Algorithm

The word frequency algorithm depends completely on the words in the corpus. Hence, it is necessary to meticulously pre-process the text data to obtain accurate results. First, all punctuation marks are removed from the text and, in case of multi-document summarization, the text is concatenated to form one whole text.

Next, stopwords are removed from the text. Stopwords are those words that do not contribute to the meaning of the passage, for example, “the”, “a”, “an”, “he”, “she”, “it” and so on.

Most times, stemming of the words is also performed. Stemming is performed using the Porter’s Stemmer method [2]. This method essential converts each variation of the word into a single root word by removing any prefixes and suffixes. For example, the words “flying” and “fly” are both reduced to “fly”.

After performing stemming, the corpus is ready for the word frequency algorithm to be applied.

5.2 TF-IDF Algorithm

This algorithm can calculate the frequency as well as the importance of a word in the document. Hence, it is not essential to remove stopwords from the corpus. On the contrary, since each word will be allocated a certain importance based on frequency, it is essential to group the similar words together.

Stemming and lemmatization is performed to the text during pre-processing. Stemming refers to reducing the word to its stem or root. Lemmatization considers the place of the word in the sentence, that is, it performs a morphological analysis of the words [10]. To understand their difference, consider the following example.

Stemming for the words “copying” and “copies” are follows:

Form	Suffix	Stem
Copies	-es	Copi
Copying	-ing	Copy

On the other hand, lemmatization for the same two words is:

Form	Morphological information	Lemma
Copies	Third person, singular, present tense of “copy”	Copy
Copying	Continuous tense of verb “copying”	Copy

From above, it can be seen that lemmatization is a more meaningful conversion of a word to its root. After applying stemming and lemmatization, the TF-IDF algorithm can be applied to the corpus.

5.3 TextRank Algorithm

The TextRank algorithm is applied to a graph in which the vertices are sentences of the corpus [3]. To generate this graph, text is first pre-processed. Steps followed in pre-processing the data are:

1. Sentence tokenization
2. Removal of punctuation marks
3. Removal of stopwords
4. Removal of duplicates
5. Changing the text to lowercase.

After the pre-processing of the text, word embeddings are used to represent the individual words are real-valued vectors. The most common algorithm used is GloVe, the Global Vectors for Word representation. It is an extension to the word2vec method of word embedding.

GloVe is an unsupervised learning algorithm used for obtaining vector representation of words [4]. It is based on two most common algorithms used: Latent Semantic Analysis (LSA) and Word2Vec.

Latent Semantic Analysis uses global statistics to derive semantic relationships between words in a corpus. The fundamental idea behind Word2Vec is that a dataset of tuples is formed consisting of (some word X, a word in the context of X). Then a neural network is applied to predict the context of X, given the word X [4]. The major disadvantage of this is that Word2Vec relies only on the local information of that language. Thus, the semantics that are learnt for a given word are only affected by the surrounding words. Word2Vec, which captures local statistics, works great with analogy tasks. On the other hand, LSA does not perform well in analogy tasks.

GloVe algorithm combines the best of both algorithms [12]. According to the author, "GloVe embeddings are a type of word embeddings that encode the co-occurrence probability ratio between two words as vector representations" [4]. GloVe uses the weighted least square objective "J". It minimizes the difference between the dot product of the vectors of the two words and the logarithm of their number of co-occurrences. The mathematical formula for "J" is as follows:

$$J = \sum_{i,j=1}^{V} f\big(X_{ij}\big)(w_i^T \overline{w}_j + b_i + \overline{b}_j - \log X_{ij})^2 \tag{8}$$

where,

w_i Word Vector of word 'i'
b_i Bias of word 'i'
$\overline{w}_j$ Context Word Vector of word 'j'
b_j Bias of word 'j'
X_{ij} number of times 'i' occurs in the context of 'j'
f is the weighted function that assigns lower weights to rare and frequent co-occurrences.

One important highlight is that the nearest neighbour algorithms can be applied to find similar words. The Euclidean distance in this space determines the similarity of two words.

Next, vectors are created for each sentence which is then plotted into a similarity matrix. A similarity matrix is a numerical representation of the similarity between any two sentences. In the paper, a cosine similarity is used to compute the similarity of two sentences [3]. This matrix is then converted into a graph where each node represents a sentence and each edge between two nodes represents their cosine similarity value. On this graph, the TextRank algorithm is applied to generate the "n" most relevant sentences for the summary.

5.4 KL Sum Algorithm

The KL Sum algorithm finds the KL divergence for a sentence to determine whether it will be included in the summary or not [11]. To calculate this divergence, again, pre-processing of text is required.

Since the divergence depends on the words in the sentence, removing stopwords in the sentence becomes essential [13]. After removing stopwords, the rest are normalized before calculating the word frequencies. In normalization, the word is converted into its base form. Normalization is usually performed in 3 steps which include: stemming, lemmatization and everything else (conversion to lowercase, removing numbers, removing punctuation and removing extra white spaces).

After performing normalization, a word frequency table is generated and term frequency of each is calculated. Then, the KL Divergence is calculated for each sentence. The sentences that have minimum divergence values are taken as part of the summary.

6 Experimental Analysis

Each of these algorithms was applied to the BBC News Dataset [14]. The dataset consists of 2225 documents from the BBC news reports corresponding to five topics from 2004 to 05. The five classes are: business, entertainment, politics, sports and technology.

The dataset also contains files of ideal extractive summaries. The similarity of the summary generated by each of these algorithms was calculated against the ideal summaries. For each of the topic areas, an average of the similarities was computed which is displayed in Table 1 (where "Ent" stands for Entertainment):

From the above table, it can be concluded that the TextRank algorithm has a higher probability of accurately ranking sentences and generating a comprehensible summary for most topics. To understand the dispersion of similarities for these algorithms, a graph was plotted as shown below.

It can be successfully concluded from Fig. 1 that although the algorithms give an average similarity index in the range of 40–60%, the spread of the similarity scores

Table 1 Average similarity index for algorithms (values are in percentages)

Topic	Business	Ent	Politics	Sport	Tech
Word Freq	49.33	48.90	51.93	**47.60**	**54.41**
TF–IDF	53.93	51.52	46.46	42.82	47.16
KL Sum	53.69	54.43	45.68	43.68	47.73
TextRank	**58.08**	**56.49**	**53.78**	44.89	51.44

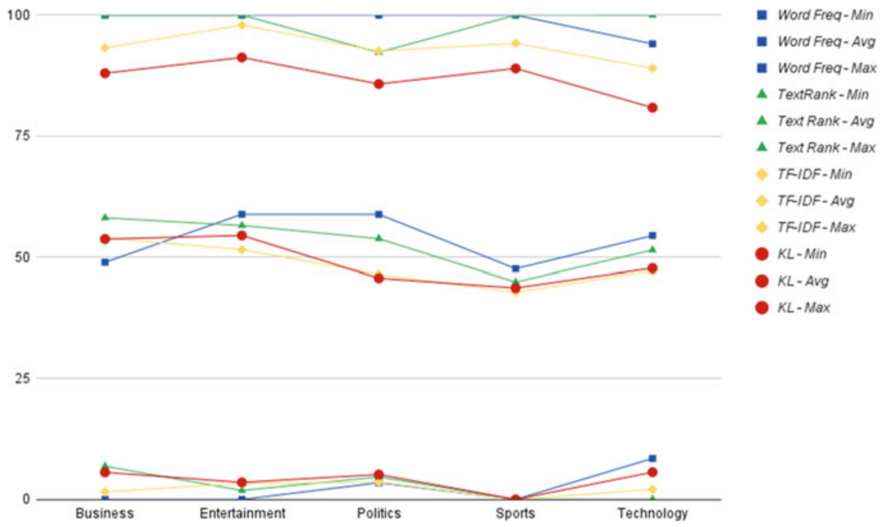


Fig. 1 Line chart showing spread of similarity values for each topic

Table 2 Standard deviation for similarity index for algorithms

Algorithm	Business	Ent	Politics	Sport	Tech
Word Freq	0.247	0.257	0.213	0.236	0.188
TF–IDF	0.186	0.184	**0.184**	0.205	0.191
KL Sum	**0.170**	**0.181**	0.190	**0.202**	**0.176**
TextRank	0.183	0.207	0.192	0.223	0.192

varies between 3 and 97%. Hence, mean of the similarity index is not an effective way to determine effectiveness of these algorithms.

Therefore, standard deviations for each of these observations was also calculated. These are tabulated in Table 2.

From Table 2, it can be inferred that in most topical areas, and the KL Sum algorithm shows minimum deviation. The algorithms that show high average similarity also exhibit higher standard deviation. This implies that although the algorithms are more accurate; there is a higher chance of inaccuracy.

Besides the mathematical differences, some other advantages and disadvantages are enlisted in Table 3.

7 Conclusion and Future Scope

Manual text summarization is an extremely tedious and time consuming task. Due to this, Automatic Text Summarizers have emerged and taken up this role.

Table 3 Advantages and disadvantages of the algorithms

Method	Advantages	Drawbacks
Word frequency	It requires less processing power. There is no need of linguistic knowledge and hence, the method is language independent	Important sentences may fail to be included in the summary due to their less score. Similarly, sentences having same meaning may be included due to high scores
TF-IDF	It is suitable for “single domain multi-document” summarization since it calculates significance as well as frequency of words	Sentences having low scores may not be included in the summary, even if they are relevant
KL Sum	It performs exceptionally well at identifying the relevance of a sentence to the topic	The summary generated is not comprehensible since the order of the sentences is vague
TextRank algorithm	It can detect coherence as well redundancy in the sentences. This algorithm is also domain-independent	TextRank algorithm disregards any word which has a lower chance of occurrence, despite maybe being meaningful in the context. This can be enhanced through the use of GloVe or Word2Vec approaches to word embeddings

There is continuous research taking place in this field, yet, it is far from being comparable to human summarization. A majority of survey papers are based on extractive summarization techniques since it is an easier approach towards the problem. Abstractive techniques require the understanding of language, like a human brain, which is a difficult task.

The main contributions of this paper include:

- Explaining the various Extractive Text Summarization methods
- Explaining the most commonly used sentence scoring algorithms for text summarization
- Providing a listing of the future research scope in this field.

The future scope of research would be to overcome the following challenges.

7.1 Challenges for Multi-document Summarization

One of the biggest challenges for multi-document summarization is redundancy [1]. There is a high chance of similar sentences being used in various documents. This may result in similar sentence scores and can affect the sentences being included in the summary. Secondly, some sentences may refer to previously mentioned content in a document. While using that sentence in the summary, context may be lost for the reader.

7.2 Challenges for Input

Presently, most summarizers work well for summarizing short length documents. When put to use on longer texts, for example, chapters of a book, the accuracy reduces significantly [1].

7.3 Challenges Related to Length of Summary

When humans summarize a text, the length of the summary varies based on the content. Deciding the stop condition for automatic summarization is challenging [1]. Setting a retention rate is the most common way, but this is not the same for all summaries.

7.4 Challenges for Evaluating the Summary

It is difficult to define what a good summary constitutes of and varies based on the type of text being summarized [1]. For example, the criteria for a good summary for sports articles and academic research articles will differ. Also, whether a summary is understood by humans or not is very subjective. Every person may select different sentences for summarization, depending on what they find useful and important. There is a need to propose methods for standardizing this.

References

1. El-Kassas WS, Salama CR, Rafea AA, Mohamed HK (2021) Automatic text summarization: a comprehensive survey. Elsevier
2. Madhuri JN, Ganesh Kumar R (2019) Extractive text summarization using sentence ranking. In: International conference on data science and communication (IconDSC)
3. Ramos J (2003) Using TF-IDF to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, vol 242, no 1
4. Shiva Kumar K, Priyanka M, Rishitha M, Divya Teja D, Madhuri N (2021) Text summarization with sentimental analysis. Int J Innovative Res Comput Sci Technol (IJIRCST) 9(4)
5. Gupta V, Lehal GS (2010) A survey of text summarization extractive techniques. J Emerg Technol Web Intell 2(3):258–268
6. Moratanch N, Chitrakala S (2017) A survey on extractive text summarization. Paper presented at the 2017 international conference on computer, communication and signal processing (ICCCSP), Chennai
7. Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Aggarwal CC, Zhai C (eds) Mining text data. Springer US, Boston, pp 43–76
8. Nazari N, Mahdavi MA (2019) A survey on automatic text summarization. J AI Data Min 7(1):121–135

9. Al-Sabahi K, Zhang Z, Long J, Alwesabi K (2018) An enhanced latent semantic analysis approach for Arabic document summarization. Arab J Sci Eng
10. Kiran Kumar G, Malathi Rani D (2021) Paragraph summarization based on word frequency using NLP techniques. In: AIP conference proceedings, vol 2317, no 1
11. Tanvi, Ghosh S, Kumar V, Jain Y, Avinash B (2019) Automatic text summarization using TextRank. Int Res J Eng Technol (IRJET)
12. Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. Stanford University
13. Baoyi W, Zhang S (2005) A novel text classification algorithm based on naïve bayes and KL-divergence. In: Sixth international conference on parallel and distributed computing applications and technologies (PDCAT'05). IEEE, 2005
14. Greene D, Cunningham P (2006) Practical solutions to the problem of diagonal dominance in Kernel document clustering. In: Proceedings of the ICML

Part II
Computer Vision

Chapter 6
Voice Activated Plant Irrigation

Akshay Tol, Prajakta Pardeshi, and Suchitra Khoje

1 Introduction

India is most popular as a country with its base as plant care, which extends to agriculture, its harvesting, and products. Indians have learnt through centuries of research, study, and experience, about how important plants are to the human life. Majority of the industries, services, products as well as innovations and inventions are established and made with plants as a source of inspiration or as the target entity. They target the growth-related factors of the plants. Out of the which, soil moisture and humidity are the most important, as they are directly related to its respiration. Household plants hold a prime place among this plant industry. Their nurture also contributes a large amount for the entire plant industry in the world. An IoT solution would be the most effective for this purpose, as it will operate on 3 levels. It will assist in real-time data gathering and presenting it. It will be low cost, requiring less power, and easily installed. And would also help in optimizing usage of labor, costs, quality, and quantity of the yield such that, water will be managed accurately during irrigation, electricity usage will be reduced, no labor will be wasted in data logging, and the timely and organized actions will result in better yields. This system is designed to automatically tackle the watering of plants as per the soil and atmospheric conditions. It is designed to be controlled using voice commands and through an Android app as well.

A. Tol (✉) · P. Pardeshi · S. Khoje
Vishwanath Karad MIT World Peace University, Pune, India

P. Pardeshi
e-mail: prajakta.pardeshi@mitwpu.edu.in

S. Khoje
e-mail: suchitra.khoje@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_6

2 Role of IoT in Agriculture

As the population growth rate increases rapidly, the agriculture industry also must improve itself in order to meet the rise in demand. To improve, it will need have to acquire the much-required edge through the adoption of new technologies. Such new technologies will not only help in improving its precision and intelligence, but also to raise its operational efficiency, lower the costs, reduce the wastage of materials, and improve the yield quality. All this is possible through IoT.

IoT-based smart farming will allow the user to keep the crops under surveillance through the means of various sensory elements that keep an eye on factors like light, temperature, soil moisture, humidity, et cetera. This type of farming is very efficient compared to the conventional view and methods.

The recent development of the drone technology can help agriculture, when incorporated, to enhance various agricultural practices. The drones can cover both sectors of a field, i.e., the ground and air, and help in the health assessment of the crops. They can also perform irrigation, spraying, and monitoring of the crops. Field and soil analysis are also possible through them.

Another possible way to utilize IoT in agriculture is to develop smart greenhouses. Different devices such as sensors, actuators, and connectivity devices can be utilized to create a controlled environment required to enhance the crop yield specifically. It totally eliminates the manual work needed normally when cultivating crops. It is a cost-effective and optimal design solution for farmers.

S. No.	Author's name	Agricultural application	Role of IoT	Use/advantage
1	Syed M. Ahmed, B. Kovela, Vinit K. Gunjan	Proposed to help farmers in cultivation of paddy, wheat, and vegetables in rural areas in obtaining yield in huge proportional	Multiple sensors to obtain soil moisture, humidity, light as well as rain, to detect automatically control irrigation. Blynk app to view sensor values	Sensors obtain data and controller constantly monitor the soil moisture, humidity, ambient light, without any physical presence needed. The farmer can view all the sensor values through the app from home
2	R. Raut, H. Varma, C. Mulla, V. R. Pawar	Proposed to help farmers check and maintain 3 major macronutrients in the soil and irrigate it	Color sensor detects nutrient deficiency in soil, while soil moisture sensor, temperature sensor, and relays irrigate the field and fertilize it	Color sensor allows to determine deficiency of which nutrient and release fertilizer soil moisture allows the controller to irrigate the field when necessary. This is communicated to the user via E-mail

(continued)

(continued)

S. No.	Author's name	Agricultural application	Role of IoT	Use/advantage
3	Coelho, Dias, Assis, Martins, Pires	Proposed to help farmers save water resources and help in making a fully automated irrigation system	Soil moisture and atmospheric sensors in a microcontroller system acquire data, sent via radio waves using LoRaWan protocol to a gateway and made available in the cloud	Soil moisture sensor senses the water content in the soil at 3 different depths, along with humidity and temperature sensors. Data are sent using LoRaWAN protocol and received using a gateway. Then, it is sent to applications or mobile devices using publisher–subscriber model. Node-RED is used as a tool in IoT-related applications that use graphical representation to display data
4	Gupta, Malhotra, Vashisht	Proposed to help farmers is severe natural disasters like draft and floods	Soil moisture and water level sensors are used to detect flood or draft conditions in the fields, hence using water suction or water irrigation pump, respectively	The sensors constantly read the water content status in the status and the water tanks. They determine whether the soil is too dry or too moist and hence either irrigation pump or suction pump is started. Also, the water level pump helps determine how much water is released in draft conditions and how much is sucked away in flood conditions. These readings are recorded in a database regularly on PC as well as in cloud. When irrigation or excess water suction is started or ended is notified to the user via alerts through SMS via GSM module
5	Monika, Yeshika, Abhishek, Sanjay, Sankar	Proposed to help farmers save water and carry out precision irrigation to yield better crops harvest	Sensors determine need for irrigating the crops. Sensor data and motor data are sent to cloud using Wi-Fi module and to users via SMS	The sensors take data such as ambient temperature, soil moisture, and light conditions and irrigate the crops properly. This data are updated into the cloud storage via Wi-Fi module, and also the user is notified using SMS via GSM module. If manual control is required, motor can be controlled using Bluetooth using a separate app

3 Literature Survey

There have been many similar projects and system attempts with variations in operating systems, controllers or processors, sensory elements, methods of irrigation, and communication protocols and types used.

Ahmed and others [1] made system with NodeMCU, sensors and Blynk app. They took input from sensors and sent them to the app. The users had a button control from app, to turn a pump on/off and thus water the plants. Raut and others [2] have used the ARM7 controller and used color, soil moisture, and temperature sensors, with output peripherals as LCD, motor drivers, and RS232 to PC system. Thus, according to sensors values, the controller fertilizes and irrigates the fields. Coelho and others [3] have PIC as the controller element, with humidity and temperature as sensing parameters, and this information is communicated to ThingSpeak, Android app, and dashboard, by sending the data using LoRaWAN technology to gateway to network server via Internet and through Node-RED to the respective destinations. Jariyayothin and team [4] have used Arduino Uno to control sensors and communicate those values to NodeMCU. NodeMCU communicates this data to a mobile app and Firebase IoT server and cloud storage. It also controls 2 solenoid valves to control the water flow from the tap and to the plants. Gupta and others [5] have used a regular Arduino Uno with multiple peripherals in their system. They take inputs from soil moisture and water level sensor and control the water suction and irrigation pumps and DC motor. If flood conditions are detected or irrigation is needed, the farmer is sent a SMS via GSM. All the readings such as sensor values and water used for irrigation are stored in a database in the form of graphs available to farmers. Suhail and others [4] have used Blynk app, IFTTT, and NodeMCU and formed a system, where an appliance is turned on or off as per a voice command, which after interpretation by IFTTT, gives an instruction to Blynk app, and thus to NodeMCU and finally to the appliance. Monica and others have used soil moisture, luminosity and moisture sensors, with relay and pump system to irrigate the plants. It can be controlled using an app via Bluetooth and stored the data on SparkFun data storage. Sensor status is also sent to user via SMS. Vaishali and others [6] have used Raspberry Pi and Android app to give warning message for pump state to user. Mahalaxmi and others [7] have used a MQTT dash app to create widgets to show latest sensor values. Premkumar and others [8] have used HTML Webpage to update pump state to user on mobile. Gulati and others [9] have used a MySQL database to store sensor data. Chidambaram and others [10] use cloud storage and Android app which can select mode show recent state of plants and set timers for watering plants. Rani and others [11] and Kumar and others [12] have used GSM to notify sensor state and the action taken to user.

3.1 Proposed System

As seen, the above systems have used many features of IoT domain efficiently. Thus, in this system, many of these features have been incorporated. This system has components like sensors which give real-time values; relay module and LED which perform the roles of driving actuators and as an alarm, respectively. The system can be controlled using Blynk app over an Internet connection using 2 buttons. Also, the sensor values can be viewed on the app through the gauge widgets on the app. The system sends sensor data to Blynk app and ThingSpeak channel through the write API ID, as soon as the data are received from the sensor. IFTTT applets are used to enable the operating of the system via voice commands. They trigger commands and response statements are defined and Web request URL as well as the pin number whose state is to be changed is mentioned. The value to be updated to the pin is mentioned, and once saved, the system responds to commands given through Google assistant.

3.2 Materials or Components Used

- **Hardware**
  - o NodeMCU
  - o Soil moisture sensor
  - o Water level sensor
  - o DHT11
  - o DC breadboard power supply module
- **Software**
  - o Arduino IDE
- **Mobile Application**
  - o Blynk
- **Cloud Storage**
  - o ThingSpeak.

4 Methodology and Results

4.1 Block Diagram

The block diagram is a combination of 3 types of blocks, such as input, processing, and output. The input blocks are sensors that react physically with the real-world

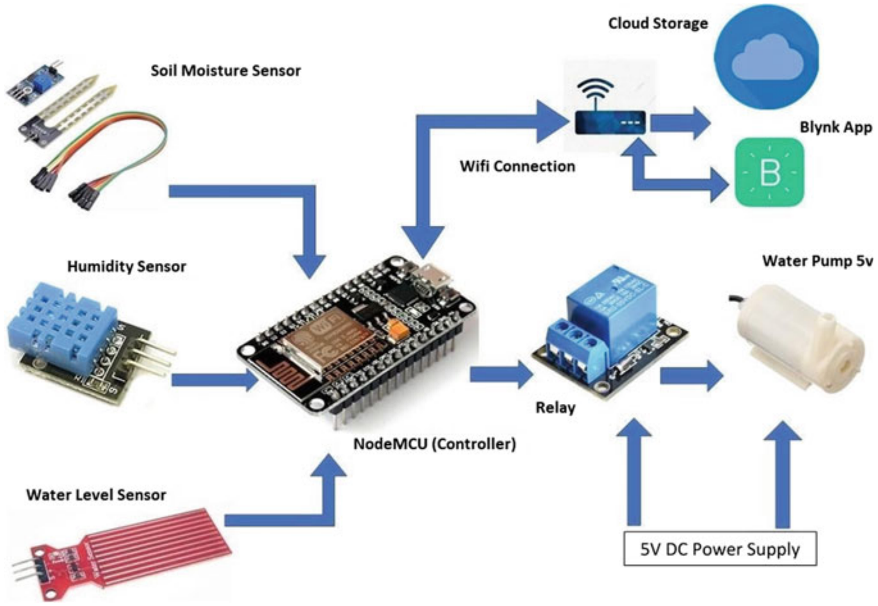


Fig. 1 Block diagram

parameters and send data to the controllers. There are 3 different sensors utilized in this system, i.e., humidity, soil moisture, and water level. The processing or controller block is a NodeMCU1.0 development board with onboard ESP8266 Wi-Fi module and 16 GPIO pins. The output block is an actuator. It can be a submersible water pump, or a solenoid valve or any mechanism that is able to switch on/off the flow of water. And to drive that actuator, a relay is required, which can switch on/off the large voltage supply. The Wi-Fi module acts a bidirectional port for sending data to cloud and receiving input from app (Fig. 1).

4.2 Hardware Setup

Below are all the modules and sensors being used in the system (Fig. 2).

Below is the completely connected hardware (Fig. 3).

4.3 Flowchart

See Fig. 4.

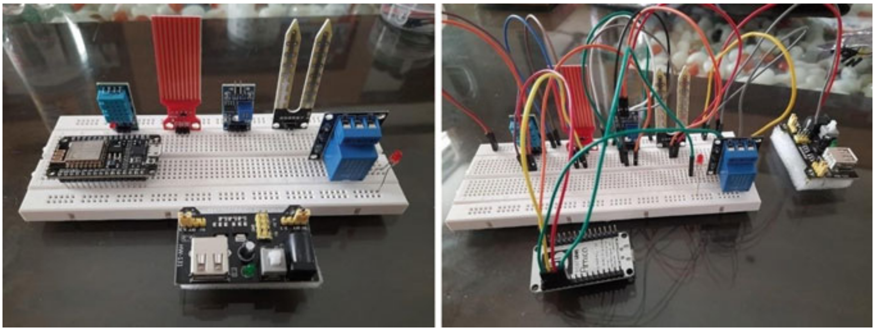


Fig. 2 Hardware components and interfacing

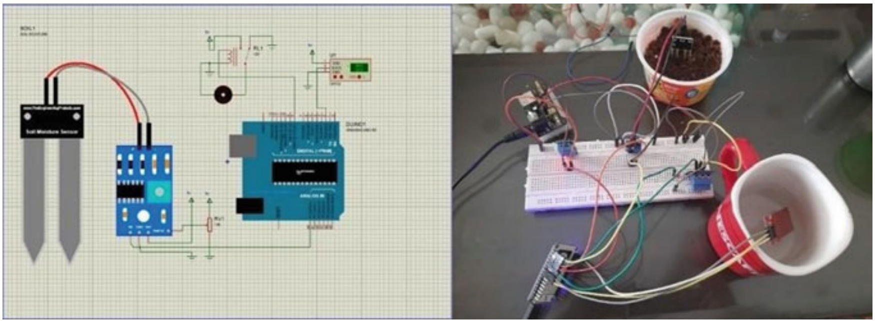


Fig. 3 Circuit schematic and complete setup

4.4 Workflow of the System

This system operates in a dual mode fashion as per the user's desire. The two modes are automatic and manual. Manual mode lets the user dictate the behavior of the system, whereas the automatic mode behaves as per the preset rules.

The system initially attempts connecting to the Wi-Fi signal. The process repeats till a connection is established properly. The system does not take any sensor values until then. After that, it attempts to connect to the Blynk server. Then, the system waits to accept any mode changes done via app. If no changes are received, the default mode is the automatic mode. Here, the NodeMCU takes readings from each sensor and sends those values to the ThingSpeak cloud server as well as Blynk app. In Blynk app, a gauge widget displays the sensor value. If manual mode is selected through the app, the relay state changes according to the pump button state changes in the app.

After the sensor readings are taken, they are converted to percentile format. These percentile values are cross referenced against predefined threshold values. For soil moisture, if the value is below 300, the relay changes state at a duty cycle of 75%.

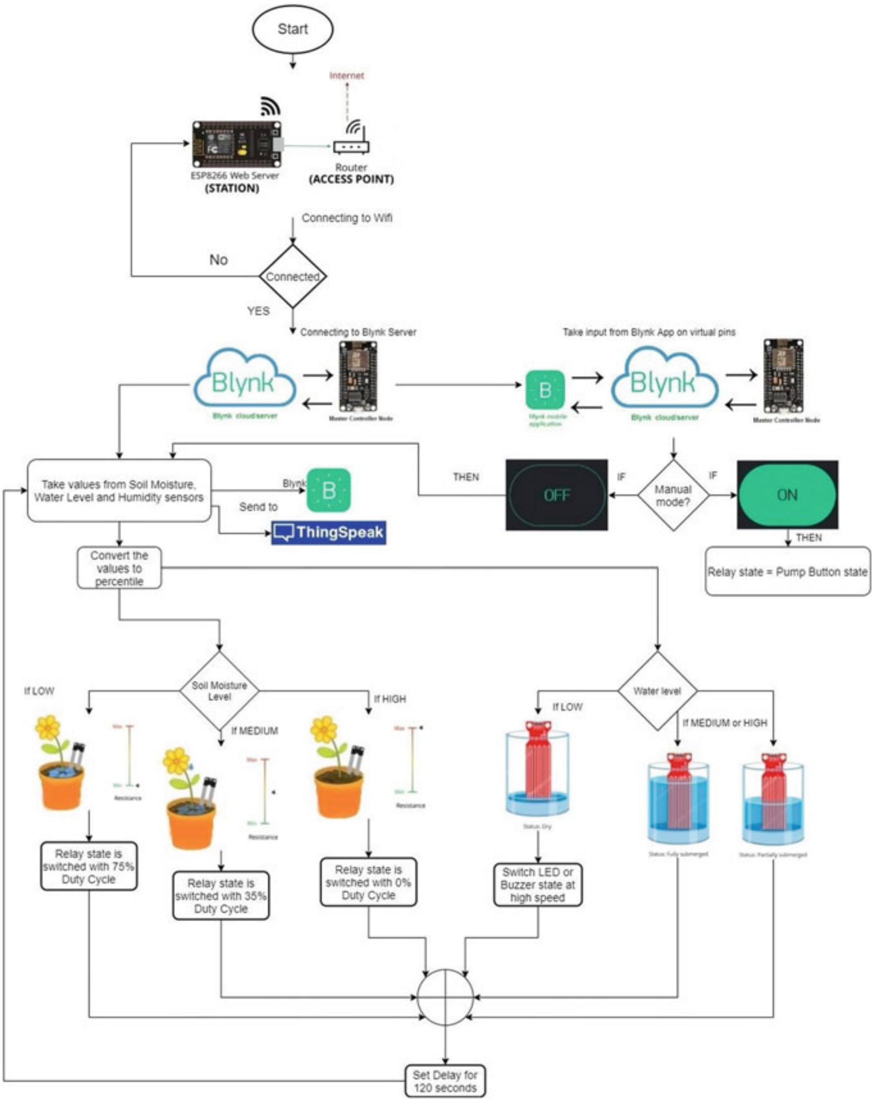

Start
Internet
ESP8266 Web Server
(STATION)
Router
(ACCESS POINT)
Connecting to Wifi
No
Connected
YES
Connecting to Blynk Server
Blynk
Blynk cloud/server
Master Controller Node
Take input from Blynk App on virtual pins
B
Blynk mobile application
Blynk
Blynk cloud/server
Master Controller Node
Take values from Soil Moisture, Water Level and Humidity sensors
Blynk
B
Send to
ThingSpeak
THEN
OFF
IF
Manual mode?
IF
ON
THEN
Relay state = Pump Button state
Convert the values to percentile
Soil Moisture Level
If LOW
If MEDIUM
If HIGH
Max
Min
Resistance
Relay state is switched with 75% Duty Cycle
Relay state is switched with 35% Duty Cycle
Relay state is switched with 0% Duty Cycle
Water level
If LOW
If MEDIUM or HIGH
Status: Dry
Status: Fully submerged
Status: Partially submerged
Switch LED or Buzzer state at high speed
Set Delay for 120 seconds

Fig. 4 Flowchart

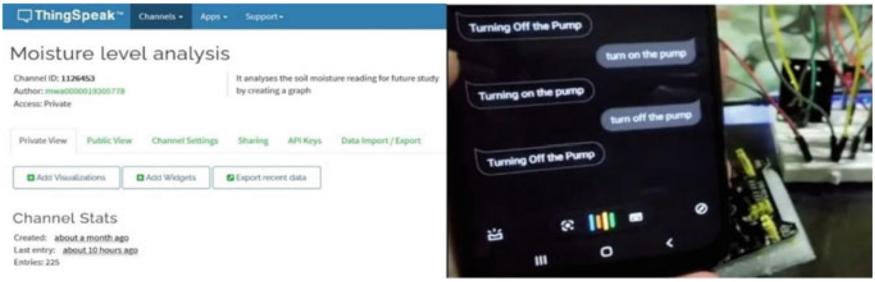
If the value is between 300 and 700, the relay changes state at a duty cycle of 35%. And if the value is above 700, the relay state stays off. These values tell if the soil has any water content or not. So, for 75% duty cycle, the soil is dry and immediate irrigation is required. For 35% duty cycle, the soil is moist and irrigation can be slow.

For water level, if the value is above 125 (a certain threshold selected), no alarm is raised. But if the value is below the said threshold, a high-speed flashing signal is applied to either a LED or a buzzer or both. The above said value indicates that the tank is about to be empty, so that it can be refilled.

After a action is taken for the taken sensor reading, the system waits for 120 s and repeats the whole process from taking a new reading from the sensors.

4.5 Blynk, ThingSpeak, and Google Assistant Output

See Figs. 5, 6, 7, and 8.



Moisture level analysis

Channel ID: 1126453
Author: mwa0000019305778
Access: Private

It analyses the soil moisture reading for future study by creating a graph

Channel Stats

Created: about a month ago
Last entry: about 10 hours ago
Entries: 225

Turning Off the Pump

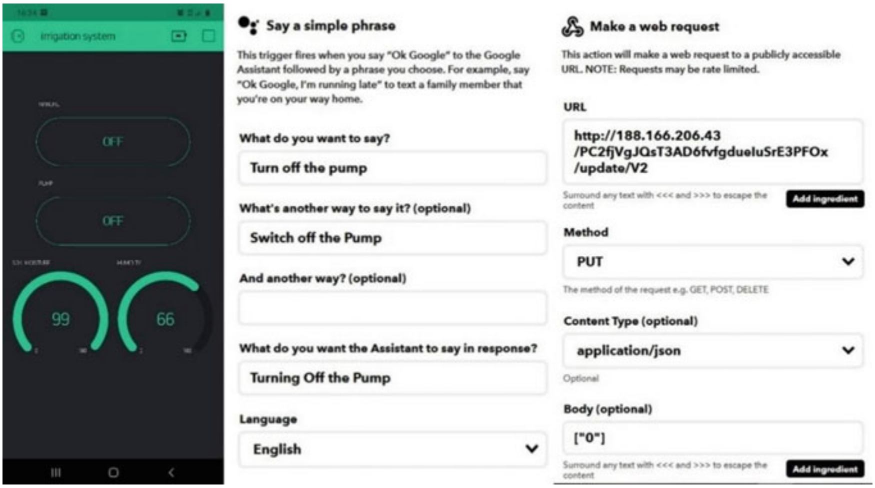
turn on the pump

Turning on the pump

turn off the pump

Turning Off the Pump

Fig. 5 ThingSpeak channel details and Google assistant working



Say a simple phrase

This trigger fires when you say "Ok Google" to the Google Assistant followed by a phrase you choose. For example, say "Ok Google, I'm running late" to text a family member that you're on your way home.

What do you want to say?
Turn off the pump

What's another way to say it? (optional)
Switch off the Pump

And another way? (optional)

What do you want the Assistant to say in response?
Turning Off the Pump

Language
English

Make a web request

This action will make a web request to a publicly accessible URL. NOTE: Requests may be rate limited.

URL
http://188.166.206.43/PC2fjVgJQsT3AD6fvfgdueluSrE3PFOx/update/V2

Surround any text with <<< and >>> to escape the content

Method
PUT
The method of the request e.g. GET, POST, DELETE

Content Type (optional)
application/json
Optional

Body (optional)
["0"]

Surround any text with <<< and >>> to escape the content

Fig. 6 Blynk app screen (left), creating an applet (mid and right)

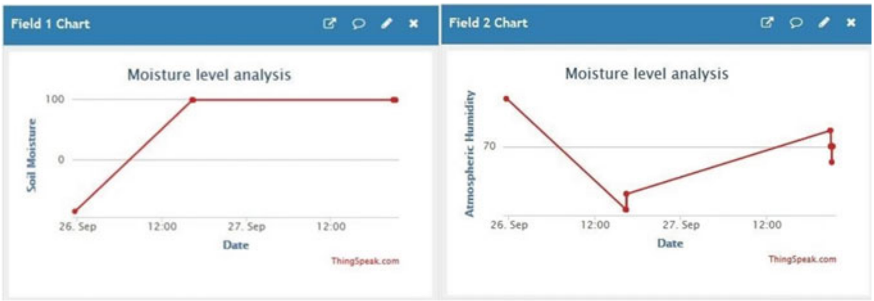
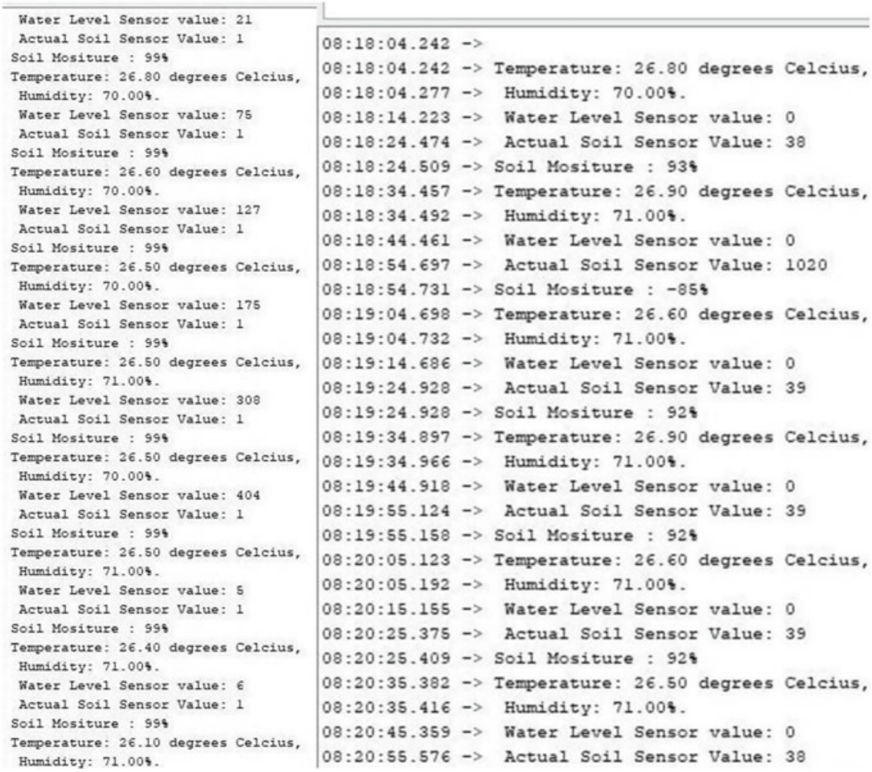


Fig. 7 ThingSpeak field graphs obtained



```
 Water Level Sensor value: 21
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.80 degrees Celcius,
 Humidity: 70.00%.
 Water Level Sensor value: 75
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.60 degrees Celcius,
 Humidity: 70.00%.
 Water Level Sensor value: 127
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.50 degrees Celcius,
 Humidity: 70.00%.
 Water Level Sensor value: 175
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.50 degrees Celcius,
 Humidity: 71.00%.
 Water Level Sensor value: 308
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.50 degrees Celcius,
 Humidity: 70.00%.
 Water Level Sensor value: 404
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.50 degrees Celcius,
 Humidity: 71.00%.
 Water Level Sensor value: 5
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.40 degrees Celcius,
 Humidity: 71.00%.
 Water Level Sensor value: 6
 Actual Soil Sensor Value: 1
Soil Mositure : 99%
Temperature: 26.10 degrees Celcius,
 Humidity: 71.00%.
```

```
08:18:04.242 ->
08:18:04.242 -> Temperature: 26.80 degrees Celcius,
08:18:04.277 ->  Humidity: 70.00%.
08:18:14.223 ->  Water Level Sensor value: 0
08:18:24.474 ->  Actual Soil Sensor Value: 38
08:18:24.509 -> Soil Mositure : 93%
08:18:34.457 -> Temperature: 26.90 degrees Celcius,
08:18:34.492 ->  Humidity: 71.00%.
08:18:44.461 ->  Water Level Sensor value: 0
08:18:54.697 ->  Actual Soil Sensor Value: 1020
08:18:54.731 -> Soil Mositure : -85%
08:19:04.698 -> Temperature: 26.60 degrees Celcius,
08:19:04.732 ->  Humidity: 71.00%.
08:19:14.686 ->  Water Level Sensor value: 0
08:19:24.928 ->  Actual Soil Sensor Value: 39
08:19:24.928 -> Soil Mositure : 92%
08:19:34.897 -> Temperature: 26.90 degrees Celcius,
08:19:34.966 ->  Humidity: 71.00%.
08:19:44.918 ->  Water Level Sensor value: 0
08:19:55.124 ->  Actual Soil Sensor Value: 39
08:19:55.158 -> Soil Mositure : 92%
08:20:05.123 -> Temperature: 26.60 degrees Celcius,
08:20:05.192 ->  Humidity: 71.00%.
08:20:15.155 ->  Water Level Sensor value: 0
08:20:25.375 ->  Actual Soil Sensor Value: 39
08:20:25.409 -> Soil Mositure : 92%
08:20:35.382 -> Temperature: 26.50 degrees Celcius,
08:20:35.416 ->  Humidity: 71.00%.
08:20:45.359 ->  Water Level Sensor value: 0
08:20:55.576 ->  Actual Soil Sensor Value: 38
```

Fig. 8 Water level sensor and soil moisture values

5 Future Scope and Complexities

5.1 Complexities Faced

1. NodeMCU has only one analog I/O pin. So, an analog I/O multiplexer needs to be configured and connected to the A0 pin of NodeMCU.
2. If the system is undergoing the process of transmitting data to ThingSpeak, the NodeMCU cannot converse with Blynk server, and hence any manual operation cannot be executed until the process terminates.
3. If the Google Assistant is not able to connect with Google, it cannot service or respond to any voice commands, and hence, remote control is not possible until connection with Google is resumed.
4. ThingSpeak cannot receive 2 sensor data values back-to-back. It requires at least 30 s of delay between the two data values.

5.2 Area of Application and Future Scope

This system can be modified to expand the scope of operation.

- It can be made self-sufficient by using solar panel for power generation.
- Rain sensor can be used to detect any sudden or external water flow onto the plants.
- Proximity or PIR sensors can be used to detect any foreign insects or objects to come in close proximity of the plant and generate alarm or alert.
- Gas sensors can be used to detect the presence of any harmful gasses in close proximity to the plant that my harm it.

6 Conclusion

The voice activated plant irrigation system has been implemented. The soil moisture and humidity sensors sense the soil moisture and humidity, respectively, and NodeMCU sends the data to Blynk app and ThingSpeak cloud storage server. NodeMCU decides the mode of operation according to the mode selected via Blynk app. Motor/relay can be operated when in manual mode via voice commands through Google Assistant. This system is ready to be implemented on a commercial scale.

References

1. Ahmed SM, Kovela B, Gunjan VK (2020) IoT based automatic plant watering system through soil moisture sensing—a technique to support farmers' cultivation in Rural India. In: Gunjan

V, Senatore S, Kumar A, Gao XZ, Merugu S (eds) Advances in cybernetics, cognition, and machine learning for communication technologies. Lecture notes in electrical engineering, vol 643. Springer, Singapore. https://doi.org/10.1007/978-981-15-3125-5_28

2. Raut R, Varma H, Mulla C, Pawar VR (2018) Soil monitoring, fertigation, and irrigation system using IoT for agricultural application. In: Hu YC, Tiwari S, Mishra K, Trivedi M (eds) Intelligent communication and computational technologies. Lecture notes in networks and systems, vol 19. Springer, Singapore. https://doi.org/10.1007/978-981-10-5523-2_7
3. Coelho D, Dias BG, de Oliveira Assis W, de Almeida Martins F, Pires RC (2020)Monitoring of soil moisture and atmospheric sensors with Internet of Things (IoT) applied in precision agriculture. In: XIV technologies applied to electronics teaching conference (TAEE), Porto, Portugal, pp 1–8.https://doi.org/10.1109/TAEE46915.2020.9163766
4. Jariyayothin P, Jeravong-aram K, Ratanachaijaroen N, Tantidham T, Intakot P (2018)IoT backyard: smart watering control system. In: Seventh ICT international student project conference (ICT-ISPC), Nakhonpathom, pp 1–6.https://doi.org/10.1109/ICT-ISPC.2018.8523856
5. Gupta S, Malhotra V, Vashisht V (2020) Water irrigation and flood prevention using IOT. In: 10th international conference on cloud computing, data science and engineering (confluence), Noida, India, pp 260–265. https://doi.org/10.1109/Confluence47617.2020.9057842
6. Monica M, Yeshika B, Abhishek GS, Sanjay HA, Dasiga S (2017) IoT based control and automation of smart irrigation system: an automated irrigation system using sensors, GSM, Bluetooth and cloud technology. In: International conference on recent innovations in signal processing and embedded systems (RISE), Bhopal, pp 601–607. https://doi.org/10.1109/RISE.2017.8378224
7. Vaishali S, Suraj S, Vignesh G, Dhivya S, Udhayakumar S (2017) Mobile integrated smart irrigation management and monitoring system using IOT. In: International conference on communication and signal processing (ICCSP), Chennai, pp 2164–2167. https://doi.org/10.1109/ICCSP.2017.8286792
8. Mahalakshmi M, Priyanka S, Rajaram SP, Rajapriya R (2018)Distant monitoring and controlling of solar driven irrigation system through IoT. In: National power engineering conference (NPEC), Madurai, pp 1–5.https://doi.org/10.1109/NPEC.2018.8476700
9. Premkumar A, Thenmozhi K, Praveenkumar P, Monishaa P, Amirtharajan R (2018) IoT assisted automatic irrigation system using wireless sensor nodes. In: International conference on computer communication and informatics (ICCCI), Coimbatore, pp 1–4. https://doi.org/10.1109/ICCCI.2018.8441209
10. Gulati A, Thakur S (2018) Smart irrigation using Internet of Things. In: 8th International conference on cloud computing, data science and engineering (Confluence), Noida, pp 819–823. https://doi.org/10.1109/CONFLUENCE.2018.8442928
11. Chidambaram RMR, Upadhyaya V (2017) Automation in drip irrigation using IOT devices. In: Fourth international conference on image information processing (ICIIP), Shimla, pp 1–5. https://doi.org/10.1109/ICIIP.2017.8313733
12. Rani GE, Deetshana S, Naidu KY, Sakthimohan M, Sarmili T (2019) Automated interactive irrigation system—IoT based approach. In: IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS). Tamilnadu, India, pp 1–4. https://doi.org/10.1109/INCOS45849.2019.8951382
13. Kumar KK, Rachamalla S, Laxmi SV, Sindhu P (2018)Most economical—IOT based smart irrigation using GSM. In: 3rd international conference on communication and electronics systems (ICCES), Coimbatore, India, pp 589–595. https://doi.org/10.1109/CESYS.2018.8724029
14. Gulati A, Thakur S (2018) Smart irrigation using Internet of Things. In: 8th International conference on cloud computing, data science and engineering (Confluence), Noida, pp 819–823. https://doi.org/10.1109/CONFLUENCE.2018.8442928

Chapter 7
A Study on Improving Banking Process for Predicting Prospective Customers of Term Deposits using Explainable Machine Learning Models

Mohd Zeeshan Khan, Sana Munquad, and Thota Sree Mallikharjuna Rao

1 Introduction

Marketing is an integral part of business and is synonymous with growth, and every business has some strategy to market their products. Marketing effectiveness is determined by many factors like audience, how effectively the message is conveyed and timing, etc. Banks employ marketing through various channels like Internet banking, emails, and phone call to name some. Thus, bank marketing can be defined as a process where the goal is to sell products like fixed deposits, recurring deposits, loans, and advances to potential customers [1]. The marketing process is highly inefficient as it is often difficult to reach the right audience. Banks have huge data about their customer which can be used to solve this problem. It is vital to adjust and implement the latest advancements in big data and machine learning in all sectors growing with profit. For achieving this, we will try to capture the same customer base as the original rule-based process and make it more efficient using machine learning models. Prediction strategies also need to evolve and make use of new marketing procedures such as reports, reviews, and data gathered from various sources. This way, the banks' management can use these predictions to develop or change their marketing strategies. To achieve the objective of providing a solution to the banks, the

M. Z. Khan
Tech Lead–UPI Design and development Organization, National Payments Corporation of India, Mumbai, India
e-mail: zeeshan.khan@npci.org.in

S. Munquad
Department of BioTechnology, National Institute of Technology, Warangal, India
e-mail: sanamo5@student.nitw.ac.in

T. S. M. Rao (✉)
School of Mathematics and Statistics, MIT World Peace University, Pune, India
e-mail: thotasree.rao@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_7

objectives of this study have been (i) to suggest a suitable machine learning model to the banks in order to provide the explainable AI-based solution for predicting prospective customers of term deposits, (ii) to provide explainable AI-based solution which banks can use to gain insights and feedback and use them to improve the banking process further with a goal of not losing even a single potential customer.

2 Related Research

There are numerous studies concerning bank and deposit marketing. Bank marketing has made use of data mining techniques for this purpose. Some researchers are coming up with using a rule-based approach to sell bank products and services [2]. On the other hand, researchers have used many different ML methods to get the best solution and achieve high accuracy to make such a campaign successful. For example, Moro, Cortez, and Laureano used the rminer package and R tool to compare the efficiency of support vector machines, decision trees, and Naive Bayes, they compared them using lift curve analysis and receiver operating characteristic (ROC) curve [3]. Similarly, Moro, Cortez, and Rita tested four models, like logistic regression, decision trees (DT), neural network (NN), and support vector machine (SVM) [4]. Evaluation of the lift cumulative curve (ALIFT) and area of the receiver operating characteristic curve (AUC) showed that best performance is achieved by neural networks, Nachev used a combination of cross-validation and multiple runs to partition the dataset into train and test sets and performance impact of designs on various neural networks [5].

Kim et al. state that [6] developed a deep convolutional neural networks (DCNNs) model for predicting if a person is likely to use the bank product. The model had the best accuracy in comparison with seven classifiers which was 76.70%, the data used for this evaluation were provided by Moro for public use. Though accuracy figures were not high but still were an improvement and were used for the marketing use case.

3 Proposed Approach and Models

When we are comparing machine learning algorithms, we need to define evaluation criteria in accordance with business requirements and goals. In this study, we have used a recall score of 0.99 as a benchmark as we wanted to capture the same customer base as the original process. This also makes it a solution that can be implemented by banks as they do not lose any customers, unlike other solutions.

The following steps will be involved in this study (Fig. 1):

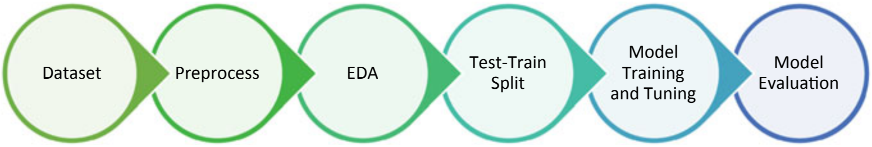


Fig. 1 Canonical steps followed by machine learning models

3.1 Dataset

Dataset chosen is from a free public dataset and client information would be retrieved using the below URL: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (Bank Telemarketing dataset: by Moro et al. [5]). This study will use the dataset available at UCI machine learning repository named, "Bank Marketing dataset." Dataset is taken from the original dataset used by the study executed by Moro et al. [5]. It has 41,188 records of the individuals with whom the telemarketing campaign was run. During the campaign stretch of May 2008–November 2010, samples were collected with 20 different fields for each person.

3.2 Dataset Preparation and EDA

In data preparation and data cleaning, data is analysed in detail for outlier detection and imputation of missing values. We will check data features for missing values and handle them accordingly, columns that have categorical values are handled by using label encoder, string, and text values will be encoded to numeric. Data transformation techniques like feature scaling standard scaling or min–max scaling will be applied.

3.3 Class Balancing

In our dataset, 41,188 instances are present out of which only 4640 have chosen to opt for term deposit giving the campaign a subscription rate of just 11.26%. So, if the model classifies all inputs as majority members will have an accuracy of 89.74%, though this model will not be of any use. Thus, the study also focuses on analyzing the implementation of various techniques to handle class imbalance using class_weights.

We have used class_weight as one of the tools to handle class imbalance, in this approach, the model is penalized for the wrong prediction made in accordance with the weight assigned to that class. The minority class is given higher weightage than the majority class, thereby preventing model bias to predict all inputs to the majority class to improve accuracy.

The following type of ML model will be used:

(a) Logistic Regression: It is a statistical approach for analyzing data with one or more explanatory/independent variables that use a binomial response variable for outcome [7, 8].

(b) K-Neighbors-Classifier: It is a non-parametric approach that assigns a data point to a class/group depending on the data points in its vicinity, k denotes the optimal value of nearby data points [9–11].

(c) Decision Tree Classifier: It is a distribution-free approach that iteratively creates a decision node that splits the record using the best attribute using attribute selection measures (ASMs). The objective is to classify the data point by learning straightforward decision rules inferred from the data features [12].

(d) Support Vector Classifier (SVC): It applies kernel function for classification, it classifies the data point by assigning it to a hyperplane which corresponds to a maximum margin between two categories [13, 14].

(e) GaussianNB: It is a straightforward yet probabilistic classifier, it presumes that the variables are highly independent. Accuracy is varying in accordance with dependency among variables [15–17].

We will be using the above five ML models for our study as they were used by the previous researchers as well but recall score which was an important metric that had taken a backseat. These five ML models are also best examples of explainable AI, therefore will be useful for business feedback.

3.4 Model Control—Hyperparameter Tuning

A hyperparameter is a characteristic of a model that is external, its value cannot be determined from the data. Hyperparameter value must be set at the start of training of the model. Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. We will use the below approach:

(i) Grid search will be used as a comprehensive search considers all parameter combinations of hyperparameters for selecting the ideal hyperparameters for a model and result in the most accurate prediction.

(ii) Randomized search CV is like grid search CV but instead of taking an exhaustive approach, it selects parameters randomly and makes combinations from all possible parameters.

(iii) The study will train the model by using the above approaches to select hyperparameter values, and for each parameter, the combination gets a score on the test set. The set of values with the best results will be selected.

(iv) The focus will be tuning the models for maximum recall performance.

Model evaluation is an important and final step in our research, for this, we will not be relying on accuracy as the only measure for model performance as this dataset is highly imbalanced. We will be evaluating the model on the following metrics:

(a) Sensitivity—This will be the prime metric as this will be of most interest to business and directly influence the success of the marketing campaign.
(b) False positive rate.
(c) ROC score.

3.5 Experiments Procedure

The individual models were trained and tested on default hyperparameters. This will help us to see what improvement is achieved at each step of our experiments and help in drawing some comparisons. The performance of different models in this step is shown in Table 1:

In terms of recall Gaussian Naïve Bayes has performed far better than others, but the classifier has a low AUC and highest FPR among all. These scores should improve after the hyperparameter tuning step has been done for each of these models.

Once hyperparameter tuning is completed, and a decision threshold is chosen to achieve desired "recall score," we will be evaluating our model on unseen data (test data) and check the model performance. In our analysis, we will check if any underfitting or overfitting is present in the model and observe if models can achieve desired recall score on our test data as well. We will be recording "recall score," "false positive rate," and "AUC score" for each model. We will be creating a confusion matrix and utilize it to get these metrics for our analysis. Below are the details for each model after hyperparameter tuning.

(a) Decision Tree: Classifier was trained using training dataset for individual models and tested on test (Table 2).

After hyperparameter tuning, recall 0.93 was recorded which was much better than our previous run with default hyperparameter values.

(b) Logistic Regression: Classifier was trained using a training dataset for individual models and tested on a test dataset (Table 3).

After hyperparameter tuning, recall 0.87 was recorded which was much better than our previous run with default hyperparameter values. This classifier was one of the fastest to train and tune.

(c) Support Vector Machine Classifier: Classifier was trained using a training dataset for individual models and tested on the test dataset (Table 4).

Table 1 Model performance without hyperparameter tuning

S. No.	ML model	AUC	FPR	Recall
1	Support vector machine classifier	0.601	0.017	0.219
2	Decision tree classifier	0.72	0.061	0.516
3	Logistic regression	0.683	0.026	0.393
4	K-neighbors-classifier	0.716	0.042	0.475
5	Gaussian Naive Bayes (GaussianNB)	0.601	0.123	0.739

Table 2 Model parameters for decision tree

Params	Tuned value	Default value
class_weight	0: 0.1 1: 0.9	0: 1 1: 1
max_depth	3	None
max_features	10	None
min_samples_split	3	2
random_state	120	None

Table 3 Model parameters for logistic regression

Params	Tuned value	Default value
class_weight	0: 0.1 1: 0.9	0: 1 1: 1
max_iter	30	100
solver	liblinear	lbfgs

Table 4 Model parameters for support vector machine

Params	Tuned value	Default value
class_weight	0: 0.1 1: 0.9	0: 1 1: 1
gamma	scale	scale
kernel	poly	rbf

After hyperparameter tuning, recall 0.86 was recorded which was much better than our previous run with default hyperparameter values. SVM took significantly more time for training and tuning than other classifiers used in this study.

(d) K-Neighbors-Classifier: Classifier was trained using a training dataset for individual models and tested on the test dataset (Table 5).

After hyperparameter tuning, recall 0.93 was recorded which was much better than our previous run with default hyperparameter values.

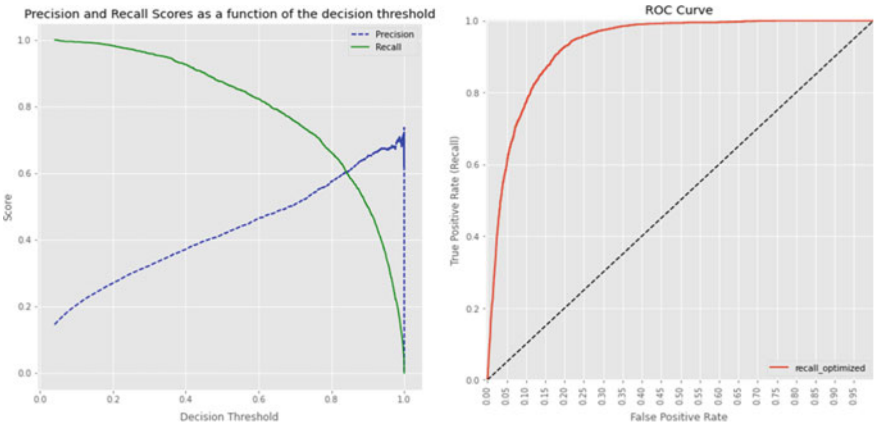
(e) Gaussian Naive Bayes (GaussianNB): Classifier was trained using a training dataset for individual models and tested on the test dataset (Table 6).

Table 5 Model parameters for K-neighbors

Params	Tuned value	Default value
leaf_size	10	30
n_neighbors	19	5

Table 6 Model parameters for GaussianNB

Params	Tuned value	Default value
var_smoothing	2.85E−07	1.00E−09



(a) Precision and recall curve (b) ROC curve

Fig. 2 Illustration: decision threshold selection for experiment 1

After hyperparameter tuning, recall 0.60 was recorded which was the same as that of our previous run with default hyperparameter values.

3.6 *Selecting Decision Threshold*

Once the ML models are tuned, we need to adjust the decision threshold for each ML model to get the benchmark 0.99 recall values. There is a definite trade-off between recall score and FPR in an ideal scenario, we should be able to have a very high recall score with very low FPR, but this is seldom the case. One must draw a balance and set the decision threshold according to business or problem requirements. Once the decision threshold is fixed can use FPR and AUC score to draw a comparison between these ML models and select the best one (Fig. 2).

3.7 *Experiments Results*

The following is the performance of selected ML models on the training set.

In Table 7, it can see that logistic regression has performed best by achieving the desired recall score with a low FPR of 0.399 which is the least among ML models which were able to achieve the benchmark for recall. Though KNN and GaussianNB had low FPR and good AUC, although they were unable to achieve a 0.99 recall score (Fig. 3).

Table 7 Training results experiment 1

S. No.	ML model	Decision threshold	Recall	FPR	AUC
1	Support vector machine classifier	−1	0.99	0.517	0.736
2	Decision tree classifier	0.1	0.984	0.477	0.753
3	Logistic regression	0.15	0.99	0.399	0.796
4	K-neighbors-classifier	0.1	0.962	0.191	0.886
5	Gaussian Naive Bayes (GaussianNB)	0.05	0.791	0.248	0.772

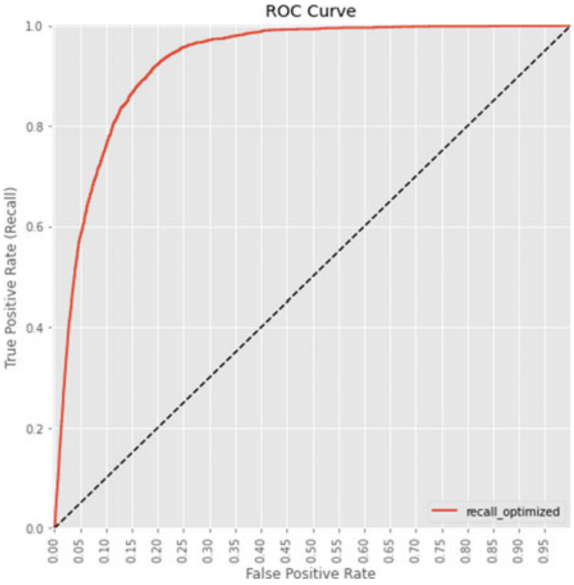


Fig. 3 ROC for logistic regression

These models had a similar performance on the test dataset and had shown no signs of underfitting or overfitting. The results on the test dataset with decision threshold set during training are listed in Table 8:

4 Results and Discussion

There were three major steps for model building. First was setting baseline performance by using untuned models for prediction, performance was not good for the SVM classifier. It had the lowest recall of 0.219, and the Gaussian Naïve Bayes classifier had the best recall of 0.739 clearly, untuned models had a bias toward the majority class and were not able to handle class imbalance.

Table 8 Testing results experiment 1

S. No.	ML model	Decision threshold	Recall	FPR	AUC
1	Support vector machine classifier	−1	0.99	0.523	0.734
2	Decision tree classifier	0.1	0.989	0.483	0.753
3	Logistic regression	0.15	0.989	0.401	0.794
4	K-neighbors-classifier	0.1	0.909	0.194	0.858
5	Gaussian Naive Bayes (GaussianNB)	0.05	0.79	0.256	0.767

Next, observations were taken after tuning these models for recall, class_weight was used as one of the hyperparameters for tuning. Results of tuning had drastically improved recall score, SVM classifier was able to achieve a recall score of 0.86, while best recall score of 0.93 was achieved by KNN classifier and decision tree classifier.

Further, for these models, decision threshold was changed to get a benchmark recall score of 0.99. It is worth noting that these models performed the same on the test dataset as they had performed on the training dataset showing no signs of underfitting or overfitting, KNN, and Gaussian Naïve Bayes classifiers were not able to achieve a recall score of 0.99. Among SVM, logistic, and decision trees classier, logistic had the best FPR of 0.401 and AUC score of 0.79 (Table 9).

We had laid out the goals and objectives through a detailed literature review of the studies done to solve this problem help us understand the existing problems and set some basic rules which helped us to make objective evaluations. It also helped us to find gaps where this study could have improved upon. Lack of baseline criteria of the evaluation was not there among different studies, some had used area of the lift cumulative curve (ALIFT), and some had used accuracy as a criterion to evaluate models. Most studies neglected the importance of customer base, to address this, we set a very high benchmark for recall score of 0.99. We also used the AUC score and FPR value in combination to benchmark recall as using recall alone we will not be able to do fair compare ML models across different experiments. This benchmark

Table 9 Results for individual ML models

S. No.	ML model	Decision threshold	Recall	FPR	AUC
1	Support vector machine classifier	−1	0.99	0.523	0.734
2	Decision tree classifier	0.1	0.989	0.483	0.753
3	Logistic regression	0.15	0.989	0.401	0.794
4	K-neighbors-classifier	0.1	0.909	0.194	0.858
5	Gaussian Naive Bayes (GaussianNB)	0.05	0.79	0.256	0.767

meant that the business could use ML models with almost no loss in customer base and still have a very efficient ML model.

5 Major Findings of the Study

This study had done a detailed literature review, where various research papers were studied and reviewed to understand their objectives and approaches.

This study identified the major gaps with the previous studies, one common gap was selecting accuracy or AUC as an evaluation metric. This caused businesses to lose a lot of potential customers and had a direct impact on their revenue. Given this understanding, we kept the recall score benchmark at 0.99, we also used the AUC score and FPR value to help evaluate our experiments to mitigate the above issue.

Another major gap was the use of complex models like ANN which provide no feedback to business, this was addressed using classical ML models for this study. This study explored traditional machine learning models which helped mitigate this issue.

6 Conclusion

The study has shown that it helps the banks to reduce customer loss to just 1% by setting a benchmark recall score of 0.99 which all models must achieve, which means that these ensemble models are more employable for business. This directly translates to a higher customer base, higher business, and higher profits. The study highlighted the effectiveness of traditional machine learning algorithms, by making the marketing process more efficient while maintaining the same results, i.e., no customer loss when compared to original rule base process. The process has become efficient by almost more than 50%.

This approach and process of evaluation can be further applied to different other use cases of marketing and similar problems where recall is of utmost importance. Further studies can utilize various other traditional and advanced ensemble ML models like stacking and blending, which may enable them to get even better values for FPR and AUC score for the same benchmark recall score or any other metrics in accordance with business requirements. Further studies can utilize various other sampling methods which were not used in this study. Building new features should be explored to find an indirect influencer that can help us get results. Other advanced ensemble approach may also be used to extend this study further and get better results.

References

1. Sing'oei L, Wang J (2013) Data mining framework for direct marketing: a case study of bank marketing. Int J Comput Sci Issues (IJCSI) 10(2 Part 2):198
2. Chun YH (2012) Monte Carlo analysis of estimation methods for the prediction of customer response patterns in direct marketing. Eur J Oper Res 217(3):673–678
3. Lawi A, Velayaty AA, Zainuddin Z (2017) On identifying potential direct marketing consumers using adaptive boosted support vector machine. In: 4th International conference on computer applications and information processing technology (CAIPT). IEEE, pp 1–4
4. Vaishnavi S, Lakshmi GD, Reddy DV, Sulochana MSN (2020) Predicting the success of bank marketing using classification techniques. J Res 6(6)
5. Moro S, Cortez P, Rita P (2015) Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst Appl 42(3):1314–1324
6. Choudhury S, Bhowal A (2015) Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. In: International conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM). IEEE, pp 89–95
7. Sperandei S (2014) Understanding logistic regression analysis. Biochemia Medica 24(1):12–18
8. Patil PS, Dharwadkar NV (2017) Analysis of banking data using machine learning. In: International conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC). IEEE, pp 876–881
9. Tomar A (2016) Various classifiers based on their accuracy for age estimation through facial features. Int Res J Eng Technol (IRJET) 3(07)
10. Zhang Z (2016) Introduction to machine learning: k-nearest neighbors. Anna Transl Med 4(11)
11. Khodabakhshi M, Fartash M (2016) Fraud detection in banking using knn (k-nearest neighbor) algorithm. In: International conference on research in science and technology
12. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
14. Das TK (2015) A customer classification prediction model based on machine learning techniques. In: International conference on applied and theoretical computing and communication technology (iCATccT). IEEE, pp 321–326
15. Bahari TF, Elayidom MS (2015) An efficient CRM-data mining framework for the prediction of customer behaviour. Procedia Comput Sci 46:725–731
16. Palaniappan S, Mustapha A, Foozy CFM, Atan R (2017) Customer profiling using classification approach for bank telemarketing. JOIV: Int J Inform Visualization 1(4–2):214–217
17. Yang FJ (2018) An implementation of Naive Bayes classifier. In: International conference on computational science and computational intelligence (CSCI). IEEE, pp 301–306

Chapter 8
Overview of Image Caption Generators and Its Applications

Shreeya Sathe, Shivani Shinde, Shriya Chorge, Shalaka Thakare, and Lalit Kulkarni

1 Introduction

Images and visuals are a dominant part of our lives and continue to be with the advent and immense use of social media. As technology evolves, people are focused on consuming it making the most of exploring how images can be perceived. A popular area where AI and deep learning help in understanding an image and generating a language description for it is Image Caption Generation.

Image Caption Generation is a tool which helps to automatically generate well-formed sentences which are concise and meaningful for a large amount of images efficiently. It not only detects objects present but also expresses all the attributes and activities present in the image. In addition to that the model requires the understanding of well-formed English phrases which is necessary to efficient visual representation and understanding (Fig. 1).

To tackle this problem, we use various deep neural network models to construct the model and handle its multidimensionality. For the first phase we use a convolutional neural network (CNN) as an image encoder which is used to extract spatial information of an input image and embed it to a fixed-length vector. Secondly, we use recurrent neural network which is a type of RNN used for generating sentences or a sequence of words in the natural language based on the image accurately by disposing of unimportant data [1].

S. Sathe · S. Shinde (✉) · S. Chorge · S. Thakare · L. Kulkarni
School of Computer Engineering and Technology, MIT World Peace University, Pune, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_8

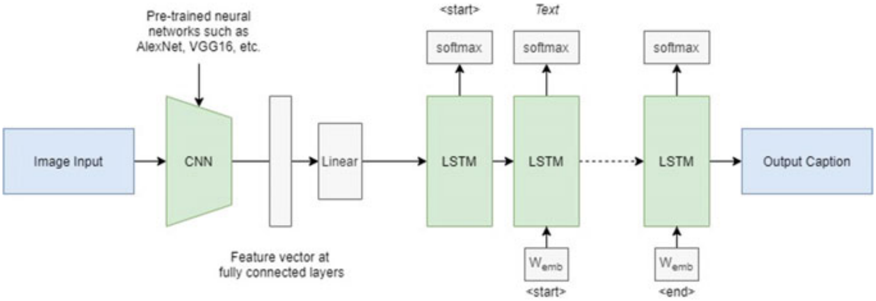


Fig. 1 Image caption generator

2 Related Work

The present solutions of Image Caption Generators are available in the following sectors.

2.1 Industrial Sector

The use of image captioning models in industrial applications will allow robots, machinery and devices to make informed decisions and perform necessary tasks independently. Image captioning enables robots to analyse and interact with their surroundings by using techniques such as object detection, face recognition, feature extraction [2].

The service robots can use the captions generated to make informed decisions. Using a template-based approach to obtain syntactically correct captions, combining existing datasets to improve object classification will ensure a more effective image captioning model and improve accuracy.

To ensure the output, that is the caption generated, is in line with the input image which contains multiple and random variables, the attention mechanism of the model should be equipped to learn the relationship between objects present in the image and the words used in the resultant caption. This method also uses a beam search decoder and the improved attention mechanism allows capturing features and higher level abstractions from the CNN [3].

As advanced Image Captioning models require compatible computing hardware which can be expensive, the method proposed in Mathur [4] uses a simple encoder and decoder with certain alterations that make it possible to implement this application using low-end hardware specifications. The model uses InceptionV4 and LSTM to encode and decode, respectively, and can be used in real-time.

2.2 Assistive Technology

Image captioning can be used to improve assistive technology and aid visually impaired to comprehend their environment. The captions generated as output can be read aloud to the users and help them to interact with it better.

Using VGG16, the project can then be incorporated in applications such as “smart glasses”, “guide dog” and “image captioning” to guide the visually impaired. Feeding the output of VGG16 to NLP, the model can generate human-like captions which will help users to get a more realistic description of their surroundings [5].

It can also be used to simplify the travelling experience by using the original Electronic Travel Aids (ETA) device suggested in Bai et al. [6]. The usage of depth and ultrasonic sensors make the device sensitive to obstacles such as transparent doors, identifying objects nearby and allow people with partial or zero visibility to move indoors almost seamlessly. Additionally, it included auditory cues to provide extra assistance to users who are completely blind.

Smartphone-based platforms can assist those with visual and disabilities by generating meaningful phrases using cameras in the device which can be narrated [7] to the users. The paper suggests a method to adopt a convenient smartphone-based platform to efficiently generate captions and text descriptions with LSTM and extracted features from VGG16.

The method proposed in Elamri and Planque [8] gives an image description using grammatically correct English phrases to help visually impaired comprehend their surroundings. Smartphones, capable of capturing photos, enable them to understand and interact with their surroundings. The CNN will help in feature extraction which can be fed to the RNN to generate a meaningful and valid description.

2.3 Agricultural Application

The method suggested in Chandeesh Kumar et al. [9] uses deep learning algorithms such as CNN and RNN to detect plants’ requirements with the help of image detection. A dataset of plants is created using a simple camera to detect the plant’s health and whether it has an adequate supply of water, nutrients, soil, sunlight, etc.

Another application in the field of agriculture would be fruit detection and segmentation. For implementing the solution, Marani et al. [10] tests four pre-trained neural networks—AlexNet, GoogLeNet, VGG16 and VGG19 to automate segmentation of grape bunches by using coloured images obtained from a RGB-D camera. The performance of neural networks was checked using two segmentation metrics, namely, segmentation accuracy and Intersection over Union (IoU). This method can be used for targeted image analysis and thus reduces manual labour.

For monitoring coffee, cocoa, tea plantations and other tree crops, information can be extracted from digital camera images and analysed. This will help farmers to monitor their crops and plantations. The poor Internet in rural areas will not pose a problem as this framework uses an on-the-go method and allows users to make decisions efficiently [11].

3 Limitations of Image Caption Generation

The techniques discussed in this paper have few limitations and are mentioned as bellow. The improvements needed to overcome these limitations are also discussed in this section.

3.1 Limitations in Industrial Applications

An industrial application of image captioning majorly discussed is that of Service Robotics [2]. A major challenge that comes up in this scenario is the sorting of relevant information that is to be used as input by the robot from the irrelevant captions generated by the system in the absence of human intuition.

Interactive machine learning can be used to enable efficient and functional training of a model [3]. However, the challenge lies in minimizing annotation effort. This can be done with the help of active learning techniques for finer training samples, in turn improving overall quality of the model.

While the model [4] does not use visual attention, adding to the overall score, it is executed at the cost of additional parameters in turn making the process of inference slower. Another challenge is the lack of standardization in splits of datasets for evaluating them offline.

In case of comparison of caption generating models using visual attention [12], the challenge lies in choosing the convolutional feature extractor. More recent architectures like Oxford VGG and GoogLeNet exhibit better performance when compared to architectures like AlexNet.

3.2 Limitations in Assistive Technology

While NLP embedded assistive technology could be a boon for the visually impaired, the results might not always be precise. Different captions based on the interpretation could be generated for a single video. In such a scenario, the generated captions may be misleading for those who are dependent solely on this assistive technology for their sense of sight.

Another challenge to consider when applying this assistive technology to real world problems is the intensive nature of computing for generating captions with respect to real-time video input. However, advances in GPU architecture and parallel computing techniques might make the timely generation of accurate captions possible in the near future.

3.3 Limitations in Agricultural Applications

The performance of the region-based approach can be enhanced by using images that focus on single objects rather than those that give a bird's eye view. Another limitation is the model can analyse only one input array at defined resolution.

For fruit detection and image segmentation, the use of a single camera will limit the perspective of the object. A one-dimensional view of the fruits or plants will not give accurate results. The small dataset of only 84 images is insufficient to train a neural network efficiently. An augmented collection of images will help yield better results and thus reduce manual labour.

The infrastructure such as cameras, sensors may not be available to farmers/smallholders in rural areas. This will complicate the process of looking after plantations and crops as they lack necessary hardware. Relying on infrastructure development in small towns, developing countries will prolong the implementation of this method.

4 Conclusion and Future Scope

In this survey, we have discussed the recent advancements in the Image Caption Generators. We have compiled several applications of Image Caption Generation, discussing how the existing technologies use this deep learning model to ameliorate their performance. Despite the remarkable growth in the number of image description systems in recent years, experimental results suggest that system performance still falls short of human performance. Nevertheless, we are aware that significant challenges still remain in various applications, and hope that the development of more powerful image caption generators will underpin further improvements to the metric.

One of the main aspects of improvement is the accuracy of the model. Better accuracy will guide users—humans, machines, robots, etc. in the decision making process. This will improve interactions between the users and their surroundings. The hardware requirements of an image caption generation system could be made more easily accessible to small towns, villages as they can implement it to improve the agricultural process. Another area of application we identified is surveillance systems which can prevent mishaps in premises by sending real-time alerts to users. We will further explore this application in our upcoming edition.

References

1. Ghoneem M, Kulkarni L (2017) An adaptive mapreduce scheduler for scalable heterogeneous systems. In: 1st international conference on data engineering and communication technology, ICDECT 2016 Proceedings, vol 2, no 469. Springer-Verlag, Germany; Lavasa City, Pune-India, pp.603–611
2. Luo RC, Hsu Y-T, Wen Y-C, Ye H-J (2019) Visual image caption generation for service robotics and industrial applications. In: IEEE international conference on industrial cyber physical systems (ICPS)
3. Biswas R, Barz M, Sonntag D (2020) Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. Springer
4. Mathur P, Gill A, Yadav A, Mishra A, Bansode NK (2017) Camera2Caption: a real-time image caption generator. In: 2017 international conference on computational intelligence in data science (ICCIDS)
5. Makav B, Kılıç V (2019) A new image captioning approach for visually impaired people. In: 11th international conference on electrical and electronics engineering (ELECO)
6. Bai J, Lian S, Liu S, Wang K, Liu D (2017) Smart guiding glasses for visually impaired people in indoor environment. In: IEEE transactions on consumer electronics, vol 63, no 3
7. Makav B, Kılıç V (2019) Smartphone-based image captioning for visually and hearing impaired. In: 2019 11th international conference on electrical and electronics engineering (ELECO)
8. Elamri C, de Planque T (2016) Automated neural image caption generator for visually impaired people. Stanford University
9. Chandeesh Kumar S, Hemalatha M, Badri Narayan S, Nandhini P (2020) Region driven remote sensing image captioning
10. Marani R, Milella A, Petitti A (2021) Deep neural networks for grape bunch segmentation in natural images from a consumer-grade camera. Precision Agric 22:387–413
11. Putra BTW, Soni P, Marhaenanto B, Pujiyanto, Fountas SSAS (2020) Using information from images for plantation monitoring: a review of solutions for smallholders. Science Direct
12. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. ICML

Chapter 9
A Vision-Based Approach for Solid Waste Materials Feature Extraction Using Deep Learning Techniques

Jyoti G. Wadmare and Sunita R. Patil

1 Introduction

Current worldwide waste generation levels are roughly 1.3 billion tonnes for each year and are relied upon to increase around 2.2 billion tones once a year by 2025. India produces 100,000 metric huge amounts of waste each day from different sources like industries, household, agriculture and fisheries, etc. The method of garbage segregation is usually done by the waste pickers which may dangerous to human health if it is not properly done and it is time consuming and fewer efficient process. This accumulation of solid wastes becoming a serious concern in an urban region and it might end in environmental pollution [1, 2]. It is really very important to have an automated AI-based sustainable solution for the waste classification. Efficient recycling of waste is economic and environmentally beneficial.

Due to exponential growth in image generation, deep learning-based models are used to solve different computer vision tasks like classification, object detection, segmentation, etc. [3].

The deep learning-based models are used to classify the solid waste materials so that the waste can be classified properly from image or video data so as to maximize the recycling. Choosing the best pre-trained model after analysing its results is the primary goal of this paper. The classification plays very important role as a primary stage for vision-based approach: Object Detection [4]. The accuracy of object detection depends on the correct classification of solid waste materials. The

J. G. Wadmare (✉)
Pillai HOC College of Engineering and Technology, University of Mumbai, Mumbai, India
e-mail: jyoti@somaiya.edu

S. R. Patil
K. J. Somaiya Institute of Engineering and Information Technology, University of Mumbai, Mumbai, India
e-mail: vice_principal@somaiya.edu

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_9

automated classification will be done by the different pre-trained models VGG16, VGG19, AlexNet and Restnet50 [5–8].

The paper is organized as follows: Section 2 gives an overview of the related work; Section 3 feature extraction using pre-trained models; Section 4 presents the experimental results on custom dataset including Trashnet dataset [9]; and paper is concluded with summary and future work.

2 Related Work

In this section, we review the approaches used for solid waste material classification and segregation.

V. P. Brintha et al have proposed automated recognition system using deep learning algorithms, i.e. Basic convolutional neural network model and region-based convolutional model to classify waste into two categories as biodegradable and non-biodegradable so as to increase the recycling rate. Dataset contains 300 images of 7 classes of waste images [10].

Sreelakshmi K et al adapted CNN-based capsule neural network for solid waste separation into plastic and non-plastic. This task of separation viewed as remarkable as per today's waste generation due to unavailibity of human labour. They have collected images of wastes from different websites [11].

Piotr Nowakowski et al proposed basic CNN is used to classify the type of E-Waste like TV, Refrigerator, TV, washing machine, etc. faster region-based convolutional neural network to detect the category of E-Waste. They have considered limited number of classes of E-wastes [12].

Olugboja Adedeji et al adapted pre-trained RestNet-50 convolutional neural network model to classify the waste materials into different categories like glass, metal, paper, etc. [13].

A V Seredkin et al proposed a method of detecting and classifying solid waste on a conveyor line using image processing and neural network. Near about 13000 municipal waste images was created to train the neural network [14].

Stephenn L. Rabano et al adapted pre-trained MobileNet model for common garbage classification. The dataset contains 2527 trash images of 6 categories [15].

Divyansh Singh et al proposes an image-based classification of polythene bags using deep learning classification model and they have presented the statistical analysis of performance on the plastic dataset which contains 400 images of the polythene class, 1200 images of the non-plastic class and 500 of the other plastic classes [16].

After extensive literature survey, different classification algorithms were used like KNN classifier, SVM classifier, Softmax classifier, fully connected neural network and basic convolutional neural networks. Because of slow error rate, convolutional neural network are best for feature extraction from images.

In this paper, we analysed the performance of pre-trained models like VGG16, VGG19, ResNet50 and AlexNet.

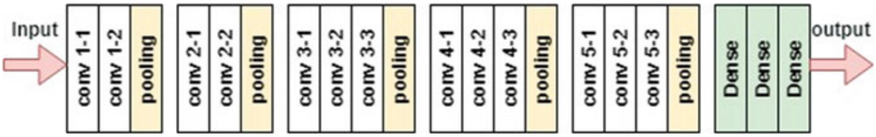


Fig. 1 VGG16 convolutional neural network model

3 Feature Extraction Using Pre-trained Models

Deep learning has many application in various fields like agriculture, medical, home automation, video surveillance, smart vehicle, face recognition, crowd detection, unmanned Arial vehicles (UAVs), drug discovery, etc. more time is needed may be days or week to train a deep learning model on big dataset.

To overcome this problem, a popular approach of transfer learning [17] where pre-trained models are used as a starting point in computer vision task, natural language processing task. It is a technique where a model trained on one task and reused it for another task. Such deep learning models pre-trained for challenging image classification problem such as ImageNet which has 1000 classes.

3.1 VGG 16 Architecture

VGG is acronym for Visual Geometry Group by Oxford University; it has total 16 layers which contains tuneable parameters. It has 13 convolution layers, 3 fully connected layers. The input to the architecture is colour image of size 224 * 224. The image is then passed through a stack of convolution layers, every convolution layer has a kernel of size 3 * 3 with stride is equal to 1 as shown in Fig. 1. Each convolution layer uses rectified linear unit RELU activation function [18]. It uses row and column padding so that size of input feature map and size of output feature map remains same. In this architecture, 5 max pool layers, the max pool operation is carried out 2 * 2 window size with stride is equal to 2.

The last fully connected layer has softmax activation function and it has 1000 channels, one for each category of image of ImageNet database.

3.2 VGG 19 Architecture

VGG19 architecture is a another variant of VGG, it has 16 convolutional layers, 3 fully connected layers, 5 max pool layers and 1 softmax layer. This architecture also requires image size (224 * 224 * 3) as input. It uses kernel size of 3 * 3 with stride is equal to 1. Max pooling and padding operations are same as VGG16 architecture.

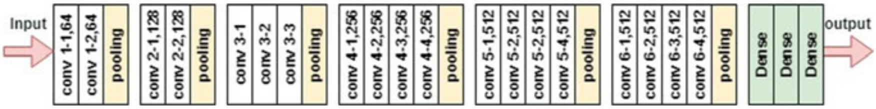


Fig. 2 VGG19 convolutional neural network model

It has 3 fully connected layers, i.e. dense layers as shown in Fig. 2. First two dense layer size is 4096 and last fully connected layer has 1000 channels.

3.3 RestNet50 Architecture

ResNet50 convolution neural network has first layer of convolution with 64 kernels of size 7 * 7 and stride is equal to 2, max pooling layer with stride is equal to 2 and then stack of convolution layers with different kernels with different sizes. The lastly average pooling and end up with fully connected layer of 1000 nodes as shown in Fig. 3.

Deep neural networks are hard to train because of vanishing gradient problem while updating the weights during back propagation. It uses the concept of skip connection means adding actual input to the output which helps to resolve the problem of vanishing gradient.

$$Y = F(X) + X \tag{1}$$

To make a function $F(X) = 0$, so that we can make output and input equal. It has total 50 layers.

The output size of image after convolution layer is calculated by

$$((n + 2p - f)/s) + 1 * ((n + 2p - f)/s) + 1 \tag{2}$$

where n is original image size, p is padding, f is filter size and s is stride.

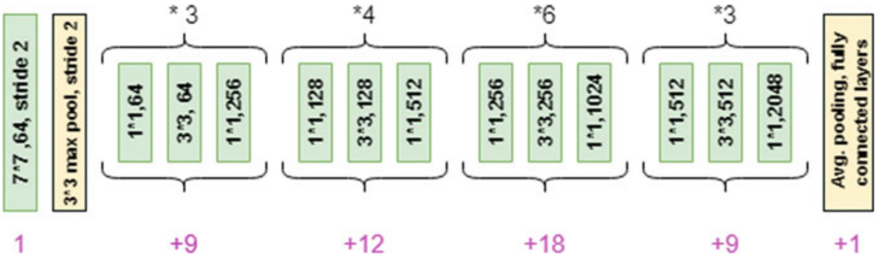


Fig. 3 ResNet50 convolutional neural network model

If input size and output size is similar then $Y = F(X) + X$ using skip connection but If input size and output size different then convolutional layer(1 * 1) is added in shortcut path or padding can be done.

3.4 AlexNet Architecture

AlexNet consist of 8 layers. Out of these 8 layers, 5 are convolution layers and 3 are fully convolution layers. The input image size it requires is 227 * 227 * 3. First convolution layer has 96 filters of size 11 * 11, stride = 4 and with RELU activation function. The output feature map after first convolution operation is 55 * 55 * 96. The channels in output feature map are equal to number of filters in the convolution layer as shown in Fig.3. The output feature map without padding operation is calculated by using following equation.

$$((n - f)/s + 1) * ((n - f)/s + 1) \qquad (3)$$

where n is image size, f is filter size and s is stride.

The output feature map with padding operation is calculated by using following equation

$$((n + 2p - f)/s) + 1 * ((n + 2p - f)/s) + 1 \qquad (4)$$

The next operation is max pooling of size 3 * 3 and stride = 2, then we get feature map 27 * 27 * 96. The next layer is convolution layer has 256 filters of size 5 * 5 with padding = 2, now the output size is 27 * 27 * 256. Likewise pooling and convolution operations are carried out for remaining layers of AlexNet as shown in Fig.4

AlexNet architecture has 60 millions of parameters, it uses two dropout layers and data augmentation operations to reduce the problem of overfitting.

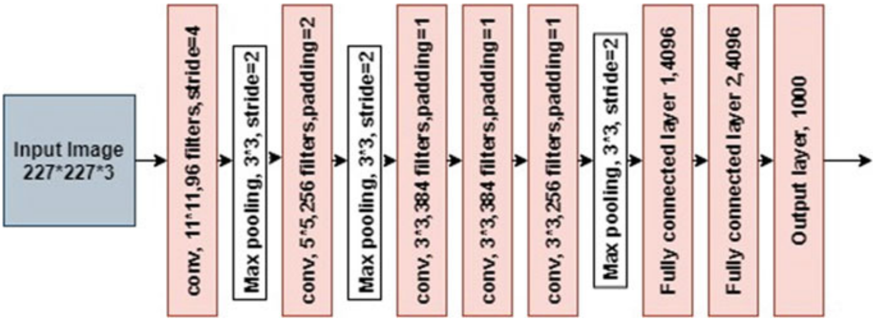

Input Image
227*227*3
conv, 11*11,96 filters,stride=4
Max pooling, 3*3, stride=2
conv, 5*5,256 filters,padding=2
Max pooling, 3*3, stride=2
conv, 3*3,384 filters,padding=1
conv, 3*3,384 filters,padding=1
conv, 3*3,256 filters,padding=1
Max pooling, 3*3, stride=2
Fully connected layer 1,4096
Fully connected layer 2,4096
Output layer, 1000

Fig. 4 AlexNet convolutional neural network model

4 Result and Discussion

All the deep learning models training were performed using Tesla K80 GPU with batch size of 32 and Epoch 20. The sample images of custom dataset contains small solid waste items of different categories like glass, metal, paper, trash, plastic is as shown in Fig. 5 and also used Trashnet dataset images.

4.1 Feature Extraction

Feature extraction is an important process before any computer vision task. The input images are passed through VGG16 convolution layers which outputs or extracts stack of visual features.

The number of filters used in first convolution layer is 64 with size 3 * 3 as shown in Fig. 6. The sample input image of glass is given to VGG16 model as shown in Fig. 7 and how features are extracted from input sample image by applying the 64 filters in first layer of convolution in VGG16 as represented in Fig. 8.The number of filters are increasing as we go to the next convolution layers as per the architecture diagram. To test the feature extraction of VGG19, Sample image of metal Fig. 9 is given as input to VGG19 and first convolution layer feature extraction is as shown in Fig. 10. Similarly, we have tested AlexNet model feature extraction of sample plastic image Fig. 11 and its output after applying 96 filters of size 11 * 11 as shown in Fig. 12

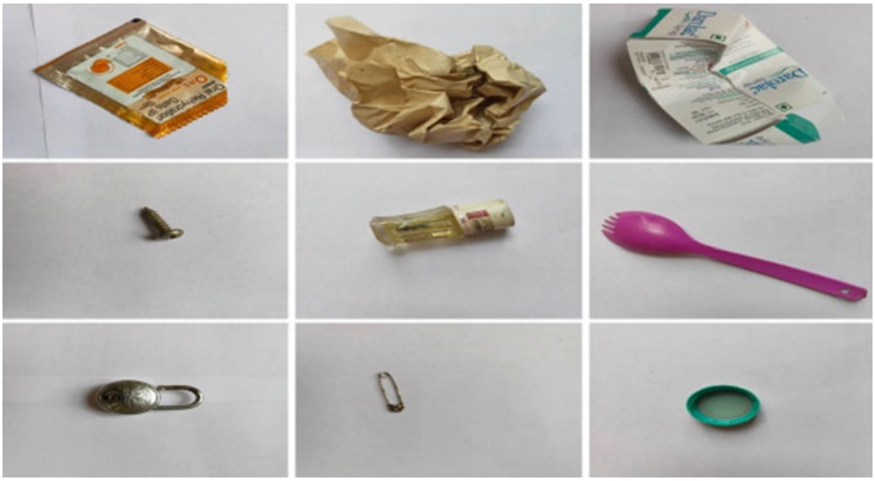


Fig. 5 Sample images of custom dataset of solid waste material

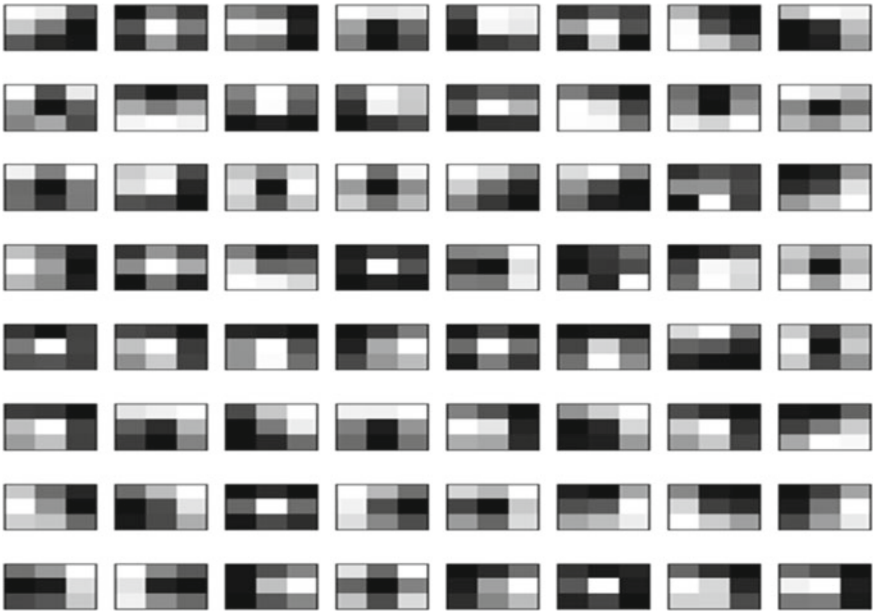


Fig. 6 First convolution layer filters of VGG16

Fig. 7 Sample image of glass

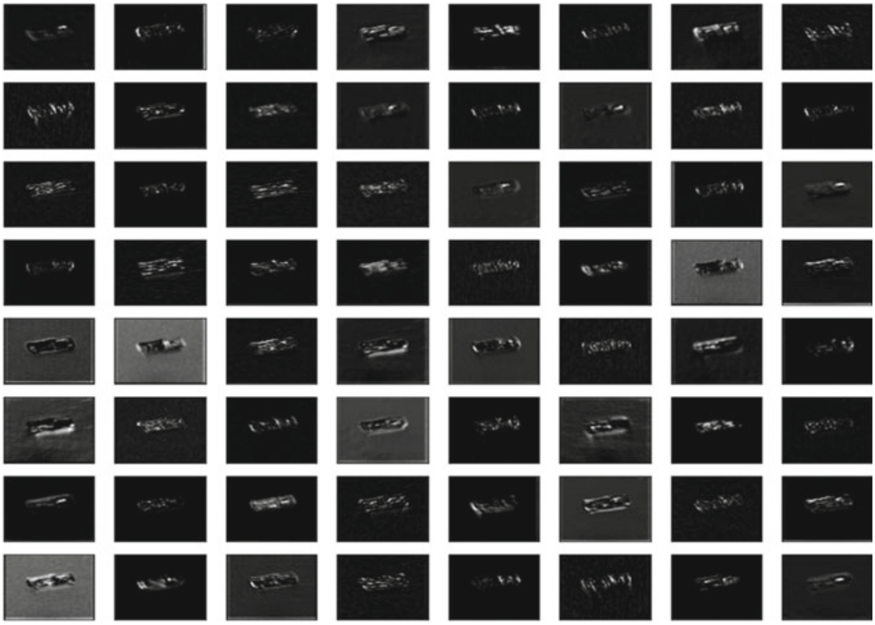


Fig. 8 Feature extraction after applying 64 filters (3 * 3 size) of first convolution layer of VGG16

Fig. 9 Sample image of metal

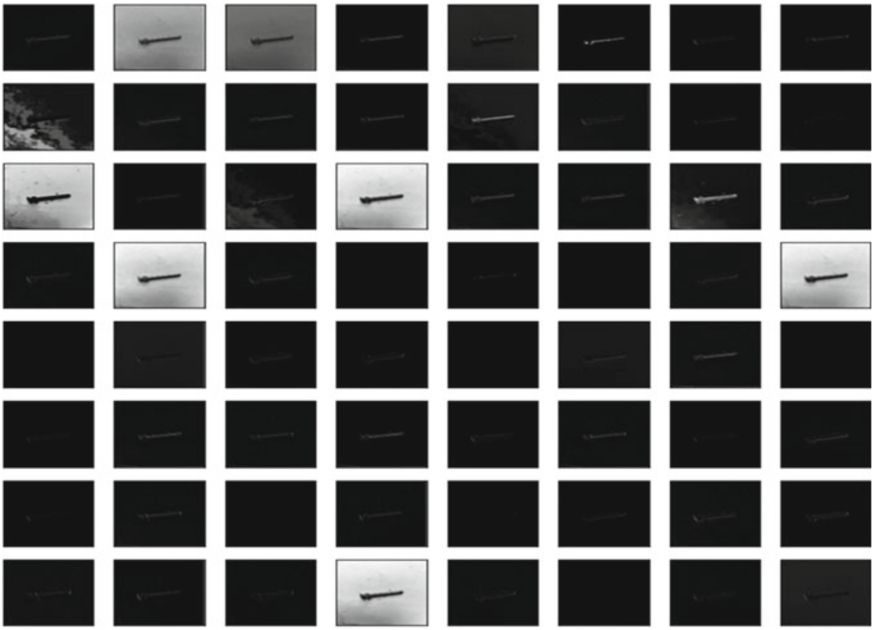


Fig. 10 Feature extraction after applying 64 filters (3 * 3 size) of first convolution layer of VGG19

Fig. 11 Sample image of plastic

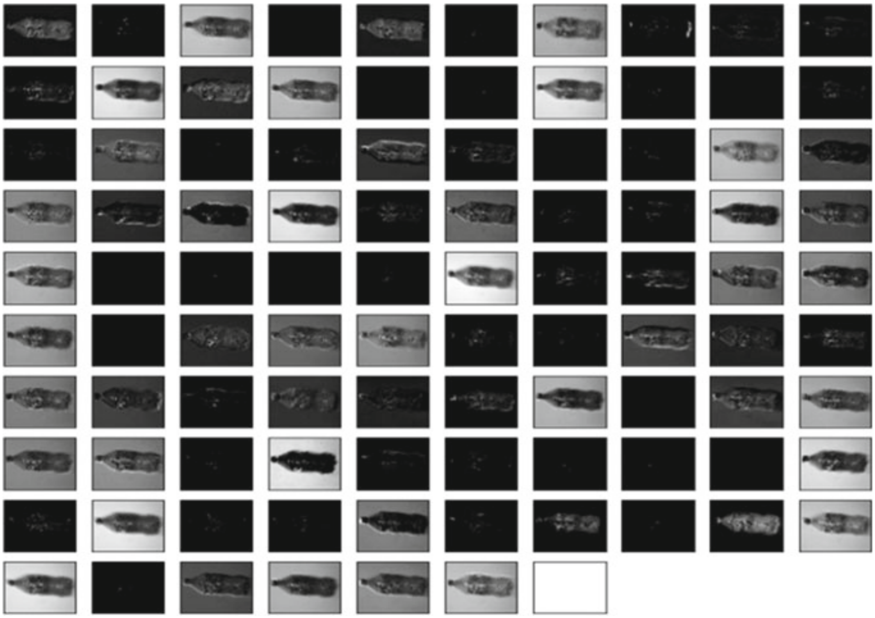


Fig. 12 Feature extraction after applying 96 filters (11 * 11) of first convolution layer of AlexNet

4.2 Deep Learning Model Training

Data augmentation was used during training of VGG16, VGG19, ResNet50 and AlexNet to avoid overfitting problem. In data augmentation, we have applied rescaling, shear, zoom and horizontal flip operation to increase the number of images during training.

In VGG 16, training accuracy was 98% and validation accuracy was 76% as shown in Fig. 13. In VGG19, Training accuracy was 96% and validation accuracy was 70% as shown in Fig. 14. In ResNet50, Training accuracy was 48% and validation

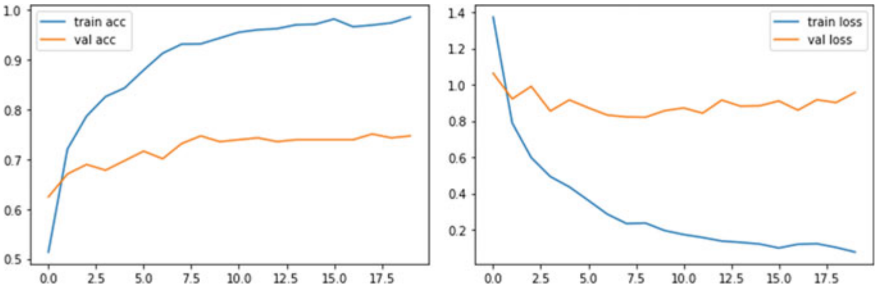


Fig. 13 Accuracy and loss plot during training and validation of VGG16 convolutional neural network model

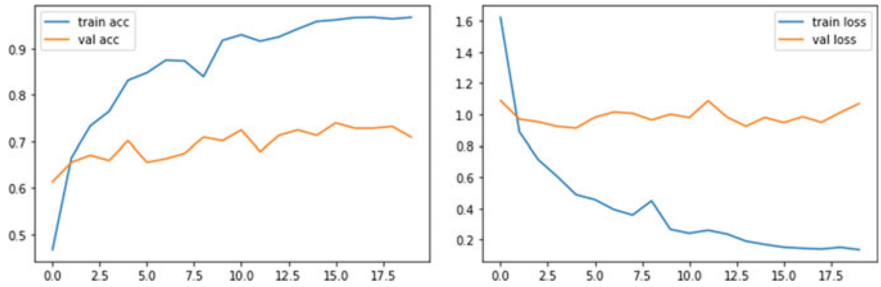


Fig. 14 Accuracy and loss plot during training and validation of VGG19 convolutional neural network model

accuracy was 39% as shown in Fig. 15. And in AlexNet, Training accuracy was 62 % and validation accuracy was 24% as shown in Fig. 16. As per results obtained, VGG16 has been accurate model as compared with other pre-trained models, where validation loss was in decreasing order and validation accuracy was in increasing order means model was working fine [19].

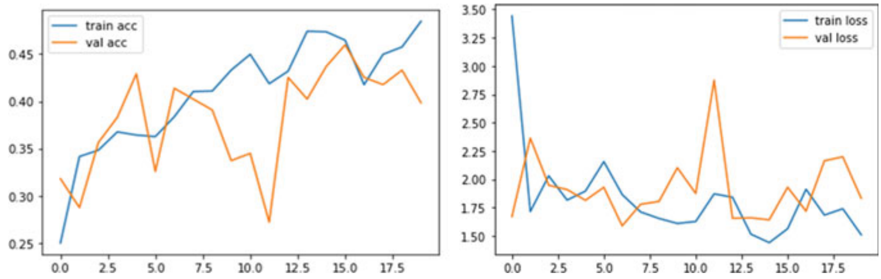


Fig. 15 Accuracy and loss plot during training and validation of ResNet50 convolutional neural network model

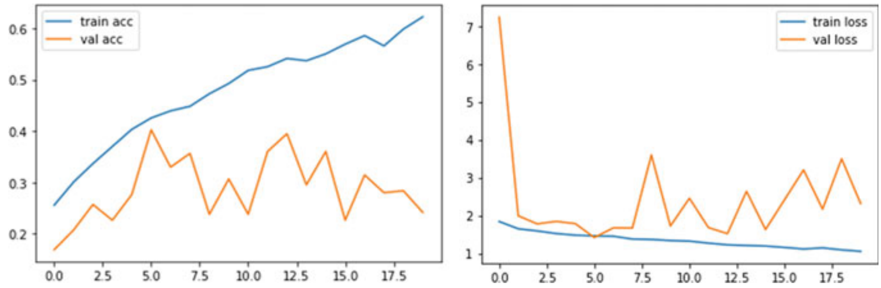


Fig. 16 Accuracy and loss plot during training and validation of AlexNet convolutional neural network model

5 Conclusion

The automated system of garbage classification is time saving and efficient method for municipal waste management. VGG16 model has been chosen for the feature extraction of input images because of its highest accuracy than VGG19, ResNet50 and AlexNet. All the models training was performed by using the dataset which comprises all kinds of recyclable waste images of different scales. During training of these pre-trained models, data augmentation process was used to increase the size of dataset and hence improve the accuracy of the model. This feature extraction task is essential step before any computer vision task such as object detection.

6 Future Work

The vision-based systems are used in the practical applications of various domains like medical, industry, agriculture, societal, etc. To provide cost effective solution of solid waste material detection from input images, the proposed approach will use weakly supervised object detection where only image level annotations are needed to train the object detection model.

Acknowledgements This work was shortlisted for the final round of Avishkar Research convention and appreciated this work in Hackathon-2021 on “Recycling Metals, Plastics & Scrap Tyres: Circular Economy Road map to Build a Green India”

References

1. Adedeji O, Wang Z et al (2019) Intelligent waste classification system using deep learning convolutional neural network. In: 2nd international conference on sustainable materials processing and manufacturing
2. Melinte DO et al (2020) Deep convolutional neural networks object detector for real-time waste identification MDPI. Appl Sci 2020(10):7301
3. Murthy CB et al (2020) Investigations of object detection in images/videos using various deep learning techniques and embedded platforms—a comprehensive review. Appl Sci 10:3280
4. Wadmare J, Patil S (2020) Improvising weakly supervised object detection (WSOD) using deep learning technique 9(3). ISSN: 2249–8958
5. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: ICLR 2015
6. Xiao J, Wang J (2020) Application of a novel and improved VGG-19 network in the detection of workers wearing masks. J Phys Conf Series, CMVIT 2020
7. Zhang KHX (2015) Deep Residual learning for image recognition, computer vision and pattern recognition, arXiv:1512.03385
8. Krizhevsky A et al (2017) ImageNet classification with deep convolutional neural networks, communications of the ACM, vol 60, no 6
9. Thung GT, GitHub repository. Available online: https://github.com/garythung/trashnet

10. Brintha VP, Rekha R, Nandhini J, Sreekaarthick N, Ishwaryaa B, Rahul R (2020) Automatic classification of solid waste using deep learning. In: Proceedings of international conference on artificial intelligence, smart grid and smart city applications
11. Sreelakshmi K, Akarsh S (2019) Capsule neural networks and visualization for segregation of plastic and non-plastic wastes. In: 5th international conference on advanced computing & communication systems (ICACCS)
12. Nowakowski P, Pamuła T (2020) Application of deep learning object classifier to improve e-waste collection planning. Elsevier
13. Adedeji O, Wang Z (2019) Intelligent waste classification system using deep learning convolutional neural network 2nd international conference on sustainable materials processing and manufacturing. Available online at www.sciencedirect.com Procedia Manufacturing
14. Seredkin AV et al (2019) Development of a method of detection and classification of waste objects on a conveyor for a robotic sorting system. J Phys: Conf Ser 1359:012127
15. Rabano SL, Cabatuan MK, Sybingco E, Dadios EP, Calilung EJ (2018) Common garbage classification using MobileNet IEEE
16. Singh D (2021) Polyth-Net: classification of polythene bags for garbage segregation using deep learning. In: International conference on sustainable energy and future electric transportation
17. Bird JJ, Faria DR (2018) A study on CNN transfer learning for image classification. In: Conference on computational intelligence
18. Wang Y, Li Y et al (2020) The influence of the activation function in a convolution neural network model of facial expression recognition. Appl Sci 10:1897
19. Salman S, Liu X (2019) Overfitting mechanism and avoidance in deep neural networks. arXiv:1901.06566v1

Chapter 10
Effective Skin Disease Detection by Analyzing Color and Texture Using Image Processing

K. Shingte and Sharmishta Desai

1 Introduction

There are so many different types of diseases, and each disease has a different impact. There are two types of diseases one is internal diseases which mostly happen because of impure blood and another are external diseases such as Acne, Eczema, Melanoma, Psoriasis, etc. A skin disease usually changes the texture and color of the skin it can also cause irritation and redness on the patient skin. Skin diseases are chronic, very infectious, and some skin diseases can develop into skin cancer if not diagnosed and treated at the right time with proper medicines. Therefore, skin diseases should be diagnosed on time to reduce its spread and seriousness. The diagnosis of skin is time consuming, and the patient has to travel all the way to the hospital; it's costly as well as is a long hectic process. Most of the common people fail to know the type of skin disease, sometime the dermatologists may also face difficulties in diagnosing the skin disease, they may require expensive laboratory tests and it's also time consuming. Skin diseases cause 1.79% of the physical disabilities of humans in the world. Around 30 to 70% people are facing skin diseases in various countries of the world. Therefore, detecting a skin disease efficiently within less time and without psychical efforts is very beneficial. In this approach, image of the infected area is taken as an input. The image undergoes feature extraction by using image processing. CNN is used for image processing and AI for classification. This approach uses algorithm like GLCM, CCM, and Stochastic Gradient Algorithm for feature extraction and prediction SVM (Simple Vector Machine) is used for image classification.

K. Shingte (✉) · S. Desai
School of Computer Science Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune 411038, India

S. Desai
e-mail: sharmishta.desai@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_10

2 Literature

Over the past few years, many researchers have performed various techniques to detect skin diseases using image processing. Image processing analysis an image to identify the type of disease. This approach is quite simple, easy, and reliable therefore many researchers have worked on implementing, few of the research papers reviewed are as follows:

Kolifi et al. [1] in this paper have briefly explained about how the system is implemented to diagnose skin diseases by using technologies like image processing and data mining. The project is divided into few parts like image processing, segmentation and feature extraction, skin disease prediction and classification model, medical treatment suggestions, or advice. Image of the infected area is taken as an input and various pre-processing techniques are applied for noise removal and image enhancement.

Utpal et al. [2] wrote about implementation of Skin Diseases Detection using Image Processing; in this paper, they've shown the implementation about detection of various types of diseases on humans, animals and as well we on plants. They have used ANN and many other pre-processing methods.

Potgieter et al. [3] in this work tell us about the implementation of the model which helps us detect the disease not just only in human beings but also in plants. Various diseases are detected using image processing, feature extraction, and segmentation.

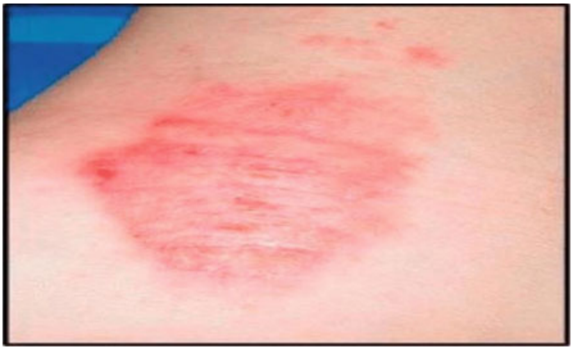
Khandelwal et al. [4] brief well about the detection of the diseases on the leaves of the plants. It uses Convolutional Neural Network (CNN), Functions like Image filtering, image compression, and image generation are used. Leaves of the plants are taken as input and then results are given.

Gaana et al. [5] researched on Diagnosis of Skin Cancer Melanoma using Machine Learning, detecting skin cancer is very expensive, lengthy procedure, and may also involve human error. Therefore, this approach was put forward by using image processing but very less complex algorithms to avoid limitations.

Bi et al. [6] discussed a state-of-the-art review on Image Synthesis with Generative Adversarial Networks; in this paper, the authors have reviewed basics of GAN's and have described some applications if image synthesis based on GANs. Al-Tuwaijari and Mohammed [7], wrote in Skin disease classification system based on machine learning technique paper about using skin biopsy, laboratory analysis, and discovery of tissue characteristics which proved effective in predicting many diseases. Namatēvs, Ivar [8] in Deep convolutional neural networks [8] discovered about theoretical and practical aspects of deep CNNs. Skin disease detection using computer vision and machine learning technique [9] by Leelavathy et al., [9], researched about different skin diseases detection using image processing methods. An analysis of convolutional neural networks for image [10] by Sharma et al., [11] focused on focuses on analyzing the performance of three popular networks: Alex Net, GoogLeNet, and ResNet50 for diagnosing skin diseases. Feature extraction and image recognition with convolutional neural networks [11] by Yu Han LIU [12], provided a technical solution for image recognition using algorithm called (CNN) which is inspired by animal visual system. Alexnet feature extraction and multi-kernel learning for object oriented classification [12] by Changmiao, Ding L, Hongyi S, Zhang W, used AlexNet features for multi-kernel learning and classification of skin diseases. Convolutional neural network based feature extraction for IRIS recognition [13] by Maram GA, Lamiaa evaluated about extracted features of a pre-trained convolutional neural network and a multi class support vector machine. Deep learning in MR image processing [14] by L D, Ko1 J, Yon J, Ryu K, this paper explanined briefly about CNN. Deep learning in MR image processing [15] by LD, K0 J, Yoon J, Nam Y implemented MR image processing and deep learning to predict skin diseases. Automating skin disease diagnosis using image classification by [16] Okuboyejo el al., [16], this paper has implemented prototyping methods to achieve detection of skin diseases. Skin disease diagnosis system using image processing and data mining by [17] Gounders et al., [17], this paper has shown the implementation of skin disease detection using image processing and data mining.

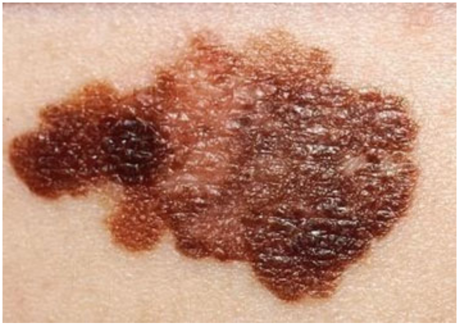
3 Data Collection

1. Eczema



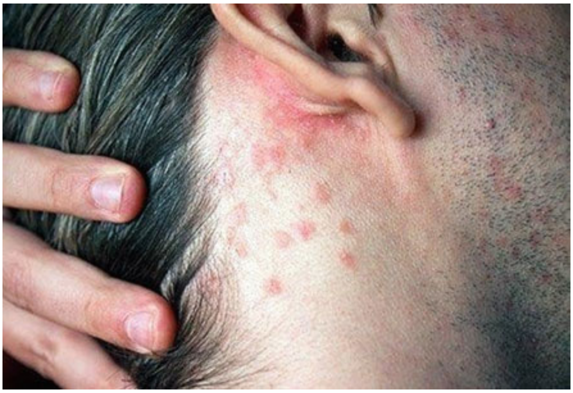
Symptoms: Dry Skin, red, itchy, and bumpy as visible in the picture.

2. Melanoma



Dangerous type of skin cancer. It causes due to melanin produced in the blood.

3. Psoriasis

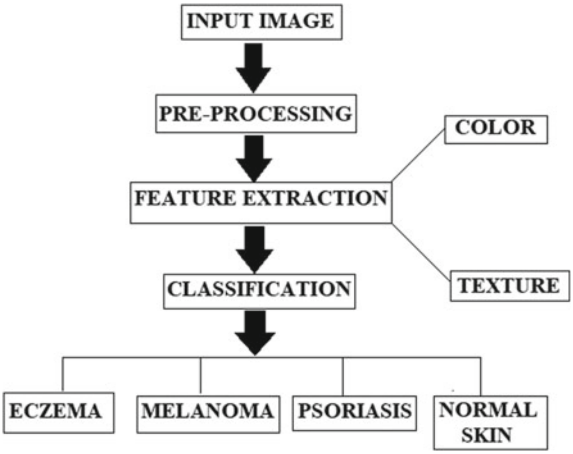


It multiples skin cells upto 10 times faster than the normal skin. The disease causes the skin to become bumpy with red patches covered with white scales.

4 Methodology

In this section, we’ll discuss steps require for implementation of the system for detection, extraction, and classification of skin diseases through analyzing the image. The architecture is split into few steps like pre-processing, feature extraction, and classification.

Have a glance at the diagram below.



4.1 Input Image

We take a digital image as an input of our implementation process. This image is analyzed for the required result.

4.2 Pre-processing

Pre-processing is the initial step of implementation after the input taken by the patient. Pre-processing is the most important step as it improves the image data that suppresses undesired distortions and enhances some image features for our further implementation. It extracts the useful information from the image. This process includes the resizing and remapping of the data if required. This technique transforms the raw data into an undesirable data.

i. Resizing—Image resizing is an important part as many times the input image is not according to the specification of the model, sometimes the image is small according to the model or sometimes it's larger. Therefore, the image needs to be resized for implementation.
ii. Remapping—It is the process of relocating the pixels. It used for lens distortion or for rotating the image if required.

4.3 Feature Extraction

It is process of extracting useful data from the input image. This process takes us close to our desired result as it gives us the most important features which help us to detect the skin disease. It's like dimensionality reductions. The input image is redundant and converted to reduced set of features. The subset of the initial features is determined. This process is called feature extraction. CNN is used for feature extraction. CNN is trained to use SGD. It uses samples from the training dataset to estimate the error gradient. The network consists of CNN, RELU, Pooling Layers, and Softmax. CNN uses Alexnet for training large neural networks quickly. It is the first CNN which uses GPU to boost performance. The methods used for extracting color and texture features are explained below:

i. Color Histogram—It the most common method which is used for extracting the color feature from an input image. It counts the similar pixels and then stores it. It represents an input from a different perspective.
ii. Gray Level Co-occurrence matrix—It is the most common method used for texture extraction from a particular image. It provides information about the texture. It is a 2nd order statistical texture analysis method.
iii. Color Co-occurrence Matrix—It also can be used to measure or extract the texture feature from the matrix.

4.4 Convolutional Neural Network

Basically, used for feature extraction. CNN is trained to use SSGD with momentum. The network consists of an input layer with pooling layers, convolutional layers, and soft max fully connected layer to extract features. CNN gives the best performance as a language processing application.

Steps of CNN:

1. Takes an input layer which is a gray scale image. Gray scale is a range of monochromatic shades that ranges from black to white. Many image editing programs such as raw.pics.io allow you to convert color images to grayscale. This removes color information. It only learns the luminance of each pixel. Black has 0 intensity and white has full intensity.
2. The output is in the form of either binary or multiclass labels.
3. Hidden layers consists of—Convolutional layer, RELU, pooling layers, and fully connected neural network.

AlexNet—CNN architecture is known as AlexNet; it was designed by Alex Krizhevsky in collaboration with Ilya Sulskever and Geoffrey Hinton. It classifies a particular input images into 1000 categories, therefore, is very rich in feature representation. It contains maxpooling layers. There are two normalization layers which are one after first and second convolutional layer. It has two layers at the top which are preceded by softmax layer. It was trained by utilizing more than 1.2 million images. We can use pre-trained networks to start our desired work. It is much more easy and robust technique. Alexnet is used in many systems such as face detection, handwriting detection image classification, text, and hypertext categorization.

5 Classification

It is a software based technology. Classification is the process of segmenting images into different categories based on their features. After feature extraction, the role of classification is to classify the pictures taken as an input; these images are taken through SVM. A SVM can prepare classifier utilizing extricated highlights from the preparing. SVM algorithm creates a line that separates the data into classes. It helps us distinguish between the different diseases.

6 Conclusion

Skin diseases are very common and thus require quick detection and treatment. Skin diseases are so dangerous that they can be life threatening if proper treatment is not given at the right time. Clinical procedure for skin detection is very expensive and time consuming. This paper explains us the implementation of a skin detection model. Image processing techniques help to detect skin diseases quickly and just by submitting an image as an input of the infected area. The extraction of image plays a very important role in image processing.

This research method was designed by using pre-trained CNN. Skin detection is very helpful for people financially and also the patients are not required to travel to the hospital for treatment. It is an easy way to know about the skin disease you're suffering with.

References

1. Kolifi AL, Enezi AL, Soliman N (2019) A method of skin disease detection using image processing and machine learning
2. Utpal S, Narang VK, Yadav N (2016) Skin diseases detection models using image processing: a survey
3. Potgieter J, Saleem M, Mahmood K (2019) Plant disease detection and classification by deep learning
4. Khandelwal S, Baranwal S, Arora A (2019) Deep learning convolutional neural network for apple leaves disease detection
5. Gaana M, Ramaiah NS, Gupta S, Diagnosis of skin cancer melanoma using machine learning
6. Bi F, Yang1 W, Yu FR, A state-of-the-art review on image synthesis with generative adversarial networks, Lei Wang, Wei Chen (Member, IEEE)
7. Al-Tuwaijari JM, Mohammed S (2020) Skin disease classification system based on machine learning technique
8. Deep convolutional neural networks (2020) Namatēvs, Ivars
9. Leelavathy S, Shobana R, Vasudevan, Prasad SS, Vasudevan JR, Prasad SS, Nihad (2020) Skin disease detection using computer vision and machine learning technique
10. Gan Q, Wei L-S, Ji T (2018) Skin disease recognition method based on image color and texture features
11. Sharma N, Mishra A, Jain V (2018) An analysis of convolutional neural networks for image
12. Yu Han LIU (2017) Feature extraction and image recognition with convolutional neural networks
13. Changmiao, Ding L, Hongyi S, Zhang W, Alexnet feature extraction and multi-kernel learning for objectoriented classification
14. Maram GA, Lamiaa, Convolutional neural network based feature extraction for IRIS recognition
15. Lee D, Ko1 J, Yoon J, Ryu K, Lee J, Nam Y, Deep learning in MR image processing
16. Okuboyejo DA, Olugbara OO, Odunaike SA (2013) Automating skin disease diagnosis using image classification
17. Gound RS, Priyanka SG, Wagh PK, Gaikwad JB (2020) Skin disease diagnosis system using image processing and data mining

Chapter 11
Construction Productivity Analysis in Construction Industry: An Indian Perspective

Poonam Katyare and Shubhalaxmi Joshi

1 Introduction

Construction industry plays a vital role in India's economy. The growth in construction industry is remarkable since from few decades. Infrastructure and social development are highly impacted by construction industry. Performance of a construction industry is depending on the successful implementation of construction projects. Smooth and timely execution of construction projects lead toward the construction productivity. Government of India provides the construction projects related to buildings for information technology due to their enormous growth at global level. Government promotes the construction projects of hospitals along with educational institutions and residential buildings.

Indian construction provides employment at a high extent. It provides goods and services to other industries. The execution of Indian construction projects is unstructured which affected by the low productivity, delays in completion of projects, unskilled employees, lack of project planning, services, and many other factors. Labors, materials, and construction equipment are the key drivers of any construction project. The efficiency of resources as construction equipment, labor and use of materials, progress in technical activities is used for construction productivity analysis of the project. Some researchers identified the factors affecting productivity of construction and analyzed it to improve the productivity. The level of productivity can be analyzed by ratio of change in total output by change in total input. Study provides insights on single factor and total factor productivity growth along with measurement parameters. Researcher also focus on the issues faced by the construction industry and provided the solution [1–3].

P. Katyare (✉) · S. Joshi
Vishwanath Karad MIT World Peace University, Pune, India

S. Joshi
e-mail: shubhalaxmi.joshi@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_11

This paper provides the review of the Indian study related to construction productivity with past specific research work provided by the researchers. Most of significant study identifies the factors or attributes with relative importance of success and failure which may affects to cost performance [4]. Some study focuses on the analysis of performances of textile industries in India and recommended supporting interrelationships of production, infrastructure, marketing, and technology and would suggest using the Policy provided by Government of India for productivity and efficiency enhancement [5]. Major contributed aspects in construction productivity consist of labor or workers working on projects, planning of projects, technical or environmental specification and equipment quality along with availability of materials in regular basis, skill of worker, motivation toward the work which lead toward the success or failure of construction project led to business development.

Many challenges faced by the Indian construction industry due to lack of unskilled employee, lack of motivation toward the tasks completion in workers, unorganized project planning, lack of technical knowledge, environmental conditions, lack of efficient equipment, involvement of huge cost related to equipment, unavailability of materials in time, lack of communication, coordination, and many other. These issues may cause the delays in project completion, increase in risk related to project and cost performance. The solution toward any one of the issues would be beneficial to the construction managers for their business development and which may affect to the growth in economy of the construction industry. Monitoring of all operations on site may help in simulation of planning and execution of the construction project. In this digital era, it is possible to keep track of the activities at construction industry.

2 Related Work

Overall construction project performance is depending on the monitoring and analysis of operations carried out at the construction site. Conventionally, analysis and monitoring of equipment and workers operations at the construction sites are addressed manually which is time consuming process and involvement of labor cost is required to execute that process which tends toward the low analysis of productivity. This results in increase in risks and issues at jobsite. Safety at jobsite is required to maintain with workers and equipment. This may take delay in completion of the projects. It is necessary to identify the factors affecting the productivity and monitor the activities at construction site.

In India, most significant study is related to the factors affecting the labor productivity and human activity identification. An assessment of productivity of labor is difficult which measures production per unit of labor working and improvement in standard of living may led to output per worker [1, 6]. Indian researchers study mostly focus on finding the critical factors affecting on cost performance of construction project as well as labor productivity of the construction project. Authors have identified critical success factors as managers skill, coordination, communication, support from management, response by the participants with environment condition. After analysis, study identified the factors affecting the cost performances of projects, are confliction between participants, illiteracy, aggressive behavior, delay in completion of the project [4]. Result indicates most significant factor with maximum positive influence on cost performance is the coordination among project participants.

Authors have done questionnaire survey-based detail analysis for finding the factors with total 112 responses using SPSS software [4]. A survey was carried out in south Gujarat, Kerala cities. For the Gujarat city 51 responses from civil contractors were analyzed through the analytic hierarchy process (AHP) and relative importance index (RII) tools and techniques. Authors identified factors from RII technique are regarding materials and labor inspiration [7]. For Kerala city 185 responses from project managers, site engineers, supervisors, and craftsmen were analyzed the critical analysis of the key factors affecting construction labor productivity by calculating the Pearson correlations among the factors/attributes. These studies are region basic and on limited dataset [8]. Some study is related to delay in completion of the construction project.

Author analyzed the construction delays variety which effect on time and cost performance. It is observed that problem of delays in construction repeated many times which may led to failure of the project and quality of the work at construction site along with dependency of one project with another project. Author has identified factors as types of delay formats with various aspects as related to production of equipment, land, unorganized pattern of working at construction site, run time withdrawal of tender, delay in execution of project. Analysis is done from published data by the flash reports with detailed information of ongoing, delayed and completed projects for the research [9]. Some study used means is a common industry standard manual for activity data analysis with single and multiple factor productivity. Increased or decreased trend of construction labor productivity is analyzed with output per labor hour [10].

The researcher suggests the area of linkage between lean construction and sustainability in the Indian construction industry through their detailed study. Authors collected the questionnaire-based data collection and analyzed data and after investigation they proposed a direct linkage between lean construction and sustainability. The authors highlight the lean tools for enabling sustainability in Indian construction industry such as: First run studies, six sigma, Kanban, last planner system (LPS) and visualization tools and identified the Just In Time (JIT) is having least impact as per RII ranking. The research study analyzed sources as management, waste reduction by considering outcome, energy minimization, elimination of non-value-added processes and health and safety improvement as the area of linkage between lean construction and sustainability in Indian construction industry [11].

Role of skill development and training are important factors for enhancing productivity and profitability in Indian construction industry. Authors found the significant relationship between these factors which shows qualitative increase in productivity [12]. Study defined the construction productivity in India along with 24 factors or attributes as base factors for the research along with two types of analysis based on the respondents which are working in Indian construction sector with survey-based questionnaire. First analysis is calculating the productivity using relative importance index and second analysis is using factor analysis/principal component analysis. As per the first analysis method, study recognized decision making, planning and logistics and supply chain management factors as most significant factors influencing on construction productivity while according to factor analysis/principal component analysis lack of planning, coordination, commitment competency, supply chain, commercial management, and inefficient site management are the seven factors affecting the Indian construction productivity [3]. Construction equipment plays an important role in construction industry. The research found on construction equipment emissions. It has been observed that despite of rigorous control and rules of construction equipment emissions in the construction, the maintenance of the equipment takes a huge cost. Research study analyzed equipment and conditions, equipment maintenance, operating conditions, and equipment operations factors which affect the rate of emissions from diesel-powered engines of construction equipment with replacement of new low-emission equipment is not required and there must be enhancement in equipment maintenance and operations which is more cost centric in reduction of emission. Study used fuel consumption factors with statistical analysis methods [13].

Indian labor productivity is analyzed by Decision Making Trial and Evaluation Laboratory (DEMATEL) with affecting factors and construction productivity analyzed by Pearson correlation analysis and linear regression analysis. Study discussed the case study and commented on the change in project affects the change in factors affecting labor productivity [14, 15] while other study recommended the factors as drawings, scope, and orders having maximum impact over the construction productivity [16]. Author used relative importance index (RII) for analysis and concluded the study with the factors affecting construction productivity are material availability, proper planning, material, resource management while factor analysis through SPSS software identified factors in different manner [17]. Some author focused on the influence of construction productivity as positive impact on construction cost performance using analytical techniques [18].

Other study discussed the factors affecting construction productivity as resource management, project management and design capability, using relative importance index scale are the important factors for construction productivity [19]. Work related to equipment in line with worker having impact on Indian construction productivity with low productivity due to factors related with manpower as well as equipment as lack of planning, lack of availability of clear work front, unskilled employee with no interest, climatic condition and unproper supervision [20]. After identification and analysis of all factors related to labor productivity it is commented that proper labor management is key factor for productivity while analyzing construction productivity management practices are influenced along with effect of age and size on the profitability and productivity of construction companies in India [21–23].

2.1 Limitation of Existing Work

Construction productivity in India is majorly concern about the factors affected by the construction projects from literature, and it is observed that Indian construction industry majorly focused on the labor productivity and the factors affecting on labor productivity [10, 14]. Existing approach for data collection is survey-based and interview-based activity which is time consuming and labor centric. It is observed that Indian construction industry is lacking behind in Automation and Digitization as compared to other countries. Region wise recognition of the factors affecting productivity and analysis of factors was implemented in India. Research study found on cost performance also but very few literatures found on construction equipment and activity recognition of all resources as workers and equipment.

3 Methodology

Our study involves review of data preparation phase along with data analysis phase.

3.1 Data Preparation Phase

Most of the study is based on the survey-based questionnaire for the data collection. A large, documented dataset is required for the research on completed construction projects. Due to lack of such dataset in India, researchers must use questionnaire survey method which covers importance of various features or parameters on project performance. Conduction of survey along with interviews with professionals from construction industry was carried out. Required modifications in the questionnaire are formulated with required parameters or factors to achieve the objective of the research study [4]. For finding the physical and financial performance of technical and productivity efficiency in textile industry data is collected from Department of Industrial Policy and Promotion, Government of India [5]. Authors used structured, close-ended questionnaire survey related on labor productivity from local industry experts, professionals, and practitioners in the Indian Construction Industry, Gujarat state of India, major cities of Kerala, sample of activities from Means manuals [3, 7–9, 11, 12, 14]. A case study-based data collection found in some work [13]. Authors used structured questionnaire survey-based methods for identifying the factors affecting the construction productivity from an interview with industry professionals [2, 17–21, 23, 24]. Other study used data of construction firms from Center for Monitoring Indian Economy (CMIE) [22].

3.2 Data Analysis Phase

Researcher used various types of analytics on data collected from survey. Parametric, non-parametric, and semi-parametric methods along with Growth Accounting Approach (GAA) which is the decomposition of change in production with respect to the change in the quantity of factors of production. Parameter free approach as Data Envelopment Approach (DEA) with multiple outputs, regression analysis, Stochastic Frontier Approach (SFA) as parametric approach with comparative analysis of results obtained from study of all approaches [1, 21]. Most of the study used multivariate technique like relative importance index (RII), reliability analysis and factor analysis for analyzing the factors used for the research [2–4, 7, 8, 11, 12, 17–19, 23, 24]. Some study focuses on the usage of data envelopment analysis technique, statistical analysis, correlation coefficient analysis, regression analysis as well as multiple linear regression analysis, Pearson correlation analysis [5, 13, 16]. Authors used SPSS 21 software tool to perform factor analysis and to validate the data reliability [17].

4 Automation in Construction Industry

There is always a challenge to monitor the real time activity and evaluation of construction resources due to the rare, dynamic, and dense nature of each construction site and operation. This skill to monitor and classify activities automatically in real time can have an advantage in achieving timely strategic operational decisions which can lead to enhanced productivity, reduced time, and cost of operations, and minimized idle times. To analyze the operations at the construction sites also control the cost of production and improve the efficiency of resources and the operational time of the resources. Study in Unites States discussed usage of sensing technologies for finding the with traditional injury prevention policies by taking inputs from project managers and safety managers throughout the Unites States which can be benefited for safety management at worksites [6]. Other study provided the usage of Global Positioning System (GPS), wireless and web-based technologies with equipment in construction sites. Collision detection and equipment tracking becomes easy by using wireless communication [25, 26]. Many of emerging technologies are available and researchers are working on machine learning approach, mobile business intelligence system, blockchain technology, video capturing at the construction worksites which led to detect an anomaly as well as construction productivity analysis. These studies majorly found in developed countries [27–38].

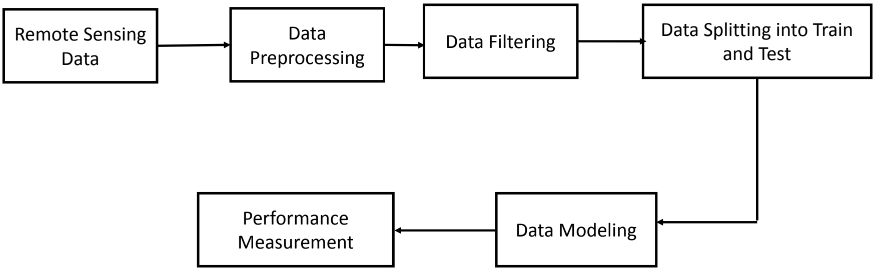


Fig. 1 Proposed Architecture

5 Proposed Work

The main objective of our study is to analyze the Indian construction productivity using advanced remote sensing data. Analyze the data by features affecting the construction productivity. Mostly, filtering methods and use of machine learning techniques and deep learning with neural network methods would help to analysis of construction productivity in Indian construction projects.

We have proposed the architecture as shown in Fig. 1.

Step 1: Data Preparation—Our study makes use of remote sensing data of construction workers or equipment. Data needs to be collected from construction industries. As per the availability of data we must apply the preprocessing of data with data cleaning and feature extraction methods. Data in appropriate format is useful for analysis purpose.

Step 2: Data Filtration—This is a step to prepare input for the modeling the data. This step would help in identification of dependent and independent variables for our study.

Step 3: Regression Analysis—This step would apply the regression model on the selected data.

Step 4: Evaluation of model—Once the model is trained, it can be tested to check the performance of the technique.

6 Conclusion

Growth in Indian construction industry highly impacted by the economy of the nation. Construction industry face the problem of low productivity, delay in completion of construction projects, cost performance, and many more factors. Most of the study in Indian construction industry is related to the analysis of factors affecting labor productivity, cost performance, equipment productivity. There is a need of monitoring the activities carried out at the construction site during the execution of the construction project regarding the workers and equipment. Tracking and monitoring of these activities may help in productivity analysis and maintaining safety, sustainability, and growth in productivity.

References

1. Basak B, Article ID : IJMET _ 08 _ 10 _ 064 in service operations : an empirical study satisfaction in service operations
2. Kathuria V, Sen K, CMDR Monograph Series No . 65 productivity measurement in Indian manufacturing : a comparison of alternative methods, no 65, pp 1–54
3. Dixit S, Pandey AK, Mandal SN, Bansal S (2017) A study of enabling factors affecting construction productivity: Indian scnerio. Int J Civ Eng Technol 8(6):741–758
4. Iyer KC, Jha KN (2005) Factors affecting cost performance: evidence from Indian construction projects. Int J Proj Manag 23(4):283–295. https://doi.org/10.1016/j.ijproman.2004.10.003
5. Bhaskaran E (2013) The productivity and technical efficiency of textile industry clusters in India. J Inst Eng Ser C 94(3):245–251. https://doi.org/10.1007/s40032-013-0073-1
6. Hallowell M, Teizer J (2010) Application of sensing technology to safety management, vol 41109. https://doi.org/10.1061/41109(373)4
7. Mistry S, Bhatt R (2013) Critical factors affecting labour productivity in construction projects: case study of South Gujarat region of India. Int J Eng Adv Technol 2:583
8. Thomas AV, Sudhakumar J (2013) Critical analysis of the key factors affecting construction labour productivity -an Indian perspective. Int J Constr Manag 13(4):103–125. https://doi.org/10.1080/15623599.2013.10878231
9. AAS (2014) Effect of construction delays on project time overrun: Indian scenario. Int J Res Eng Technol 3(1):543–547. https://doi.org/10.15623/ijret.2014.0301091
10. Vereen SC, Rasdorf W, Hummer JE (2016) Development and comparative analysis of construction industry labor productivity metrics. J Constr Eng Manag 142(7):04016020. https://doi.org/10.1061/(asce)co.1943-7862.0001112
11. Dixit S, Mandal SN, Sawhney A, Singh S (2017) Area of linkage between lean construction and sustainability in indian construction industry. Int J Civ Eng Technol 8(8):623–636
12. Dixit S, Mandal SN, Sawhney A, Singh S (2017) Relationship between skill development and productivity in construction sector: a literature review. Int J Civ Eng Technol 8(8):649–665
13. Fan H (2017) A critical review and analysis of construction equipment emission factors. Procedia Eng 196(June):351–358. https://doi.org/10.1016/j.proeng.2017.07.210
14. Chaturvedi S, Thakkar JJ, Shankar R (2018) Labor productivity in the construction industry: an evaluation framework for causal relationships. Benchmarking 25(1):334–356. https://doi.org/10.1108/BIJ-11-2016-0171
15. Dixit S, Mandal SN, Thanikal JV, Saurabh K (2019) Evolution of studies in construction productivity: a systematic literature review (2006–2017). Ain Shams Eng J 10(3):555–564. https://doi.org/10.1016/j.asej.2018.10.010
16. Dixit S (2018) Analysing enabling factors affecting the on-site productivity in indian construction industry. Period Polytech Archit 49(2):185–193. https://doi.org/10.3311/ppar.12710
17. Dixit S, Mandal SN, Thanikal JV, Saurabh K (2018) Critical analysis of factors affecting the on-site productivity in Indian construction industry, no. July, pp. 38–45. https://doi.org/10.3311/ccc2018-006
18. Dixit S, Saurabh K (2019) Impact of construction productivity attributes over construction project performance in Indian construction projects. Period Polytech Archit 50(1):89–96. https://doi.org/10.3311/ppar.12711
19. Dixit S, Mandal SN, Thanikal JV, Saurabh K (2019) Study of significant factors affecting construction productivity using relative importance index in Indian construction industry. In: E3S web conference, vol 140. https://doi.org/10.1051/e3sconf/201914009010

20. Natarajan SP (2019) Improvement of manpower and equipment productivity in Indian construction projects. Int J Appl Eng Res 14(2): 404–409. [Online] Available: http://www.ripublication.com
21. Agrawal A, Halder S (2020) Identifying factors affecting construction labour productivity in India and measures to improve productivity. Asian J Civ Eng 21(4):569–579. https://doi.org/10.1007/s42107-019-00212-3
22. Cyril EJ, Singla HK (2021) The mediating effect of productivity on profitability in Indian construction firms. J Adv Manag Res 18(1):152–169. https://doi.org/10.1108/JAMR-05-2020-0092
23. Dixit S (2021) Impact of management practices on construction productivity in Indian building construction projects: an empirical study. Organ Technol Manag Constr 13(1):2383–2390. https://doi.org/10.2478/otmcj-2021-0007
24. Ranjithapriya R, Arulselvan S (2020) Study on factors affecting equipment management and its effect on productivity in building construction. Int J Eng Res V9(04):223–230. https://doi.org/10.17577/ijertv9is040176
25. Oloufa AA, Ikeda M, Oda H (2003) Situational awareness of construction equipment using GPS, wireless and web technologies, vol 12, pp 737–748. https://doi.org/10.1016/S0926-5805(03)00057-8
26. Harichandran A, Raphael B, Mukherjee A (2020) A robust framework for identifying automated construction operations. In: Proceeding 37th international symposium automation robotic construction. https://doi.org/10.22260/isarc2020/0066
27. Jidiga GR, Sammulal P (2015) Anomaly detection using machine learning with a case study. In: Proceeding 2014 IEEE international conference advance communication control computer technology ICACCCT 2014, vol 2, n. 978, pp 1060–1065. https://doi.org/10.1109/ICACCCT.2014.7019260
28. Djatna T, Munichputranto F (2015) An analysis and design of mobile business intelligence system for productivity measurement and evaluation in tire curing production line. Procedia Manuf 4(Iess):438–444. https://doi.org/10.1016/j.promfg.2015.11.060
29. Dadhich S, Bodin U, Sandin F, Andersson U (2016) Machine learning approach to automatic bucket loading. In: 24th Mediterranean conference control automation MED 2016, pp 1260–1265. https://doi.org/10.1109/MED.2016.7535925
30. Zdravevski E et al. (2017) Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering. IEEE Access 5(c):5262–5280. https://doi.org/10.1109/ACCESS.2017.2684913
31. Xiao B, Zhu Z (2018) Two-dimensional visual tracking in construction scenarios: a comparative study. J Comput Civ Eng 32(3):04018006. https://doi.org/10.1061/(asce)cp.1943-5487.0000738
32. Seong H, Son H, Kim C (2018) A comparative study of machine learning classification for color-based safety vest detection on construction-site images. KSCE J Civ Eng 22(11):4254–4262. https://doi.org/10.1007/s12205-017-1730-3
33. Kim H, Kim H, Hong YW, Byun H (2018) Detecting construction equipment using a region-based fully convolutional network and transfer learning. J Comput Civ Eng 32(2):04017082. https://doi.org/10.1061/(asce)cp.1943-5487.0000731
34. Harichandran A, Raphael B, Varghese K (2019) Inferring construction activities from structural responses using support vector machines. In: ISARC 2018—35th international symposium automation robotic construction international AEC/FM hackathon future building things, no. July 2019, 2018. https://doi.org/10.22260/isarc2018/0047
35. Harichandran A, Raphael B, Mukherjee A (2019) Determination of automated construction operations from sensor data using machine learning. In: Proceeding 4th international conference civil building engineering informatics, pp 77–84
36. Perera S, Nanayakkara S, Rodrigo MNN, Senaratne S, Weinand R (2020) Blockchain technology: is it hype or real in the construction industry? J Ind Inf Integr 17(January):100125. https://doi.org/10.1016/j.jii.2020.100125

37. Sharma G, Kotia A, Ghosh SK, Rana PS, Bawa S, Ali MKA (2020) Kinematic viscosity prediction of nanolubricants employed in heavy earth moving machinery using machine learning techniques. Int J Precis Eng Manuf 21(10):1921–1932. https://doi.org/10.1007/s12541-020-00379-9

38. Gondia A, Siam A, El- W, Nassar AH (2020) Machine learning algorithms for construction projects delay risk prediction. J Constr Eng Manag 146(1):04019085. https://doi.org/10.1061/(asce)co.1943-7862.0001736

Part III
Communication and Networking

Chapter 12
Checkpoint-Based Blockchain Approach for Securing Online Transaction

Priyanka A. Chorey

1 Introduction

The growth of Company 4.0 and its subsequent industrialization would promote the IIoT, a system-wide data collection and analysis that will allow quicker, increasingly cost-effective, complex, and reliable operations to manufacture better value products. IIoT is among the prominent critical innovations to support intelligent industrial enterprises and markets, i.e., improving service delivery, improving production performance, allowing for preventive analytics, and fostering environmentally sustainable practices across various usages [1].

Since another day of bank crime, the approach to attain repercussions in authorities, corporate organizations, and accounting might be a widespread problem. A firm reliance on online technologies has increased bank transfers in the current world. Financial theft, nevertheless, was collectively facilitated via the trade physical and digital. As transfers became a popular form of transport, a modern methodology was already focused on dealing with the drawback of theft. Several fraudulent activities and software applications deter theft in credit cards, banking, e-commerce, insurers, and industry. There is no way to be still sure of the reality of activity or payment. The most successful probability is to track down proof of wrongdoing through available information utilizing complex equations. Big data means the complete and deductible details and facts beneficial to potential references are removed and analyzed [2].

The content being used in data analysis is derived from various data such as tenacious, chaotic, and non-systematic data carried out in the data analysis. Bank face challenges in the protection of Internet-based finance. However, many consumers were deeply frustrated by an endless amount of keys, various device symbol systems, and resources beyond the spectrum used by institutions for interaction. While

P. A. Chorey (✉)
Research Scholar, G. H. Raisoni University, Amravati, India

Assistant Professor, PRMIT & R, Badnera, Amravati, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_12

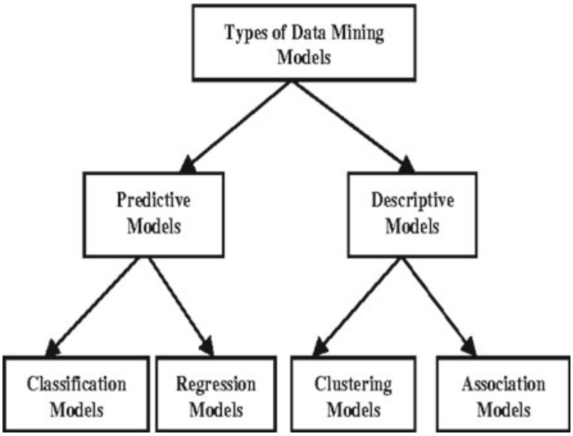


Fig. 1 Data mining models

bankers' defense becomes less vigilant to consumer convenience, affiliate development will continue to occur. But all the objects use nuanced predictive concepts and laws to evaluate suspected activity accurately. Even so, consumers are deeply upset whether they are struggling with the declining significance of the institutions because there are various threatening comments, those who should still defend bankers against everybody. Verification was once a simple guiding principle because all the increasingly daunting observation is sufficiently qualified through all ages, from watching the figures to mainly inquiries about documentation or even checking payments because of all the symbol systems and advanced contact networks. Increase the encryption to interrupt the user with no trouble, resulting in lower threat and a consumer's satisfaction value [3].

The methodology for detailed data analysis identifies the relationships or links throughout the information and determines all assets and relationships between the records/knowledge analyzed. The Reserve Bank of India (RBI) retains records on challenging extraordinary situations under the presumption of all these cases. The performing analysis under these scams is deposit-saving A/C, home mortgages, credit card and personal loans, cash loans, cash transfers, demand draft, online payments.

Throughout the financial industry, data-gathering technologies and frameworks could be used primarily for fraudulent financial reporting (Fig. 1).

1.1 Fraud Detection

Deceitful behaviors could be observed and documented and use all kinds of concepts and methods of data analysis. However, there are many techniques, whereby mining techniques are used to identify trends of corruption. It should tackle various data holdings containing financial data, using their data mining codes to assess the abuse in their data holdings. The whole initial step is using their knowledge around how deception occurs, and they may respond to such phenomena and determine the level of problem in them. In the following step, the fraudulent trend is determined based on the bankers' internal private data. Many banks just use a "hybrid" sensor to estimate theft. Data analysis is not just an essential consideration for developing new clients in financial institutions; it could also help maintain their current clients. Ratification and however doing to customers are critical problems for every company. Managers inside the financial institutions must bear in mind that the bank staff recognizes each client. The data analysis often facilitates the labeling of potential customers for products and services or the determination or exploration of the valuable purchasing habits of a customer in a way that allows banks to retain old customers prepared while offering personal motives tailored to the needs of any customer.

The ledger platform has been developed to increase the reliability of the transfer. Cryptocurrency is indeed a possible alternative for transparency of files, using encryption to avoid flaws. In certain instances, blockchain is used for encrypted connectivity and security issues. The blockchain is decentralized, and a central agency cannot authorize transfers. All service providers should agree to verify the transactions safely, and prior documents could not be reversed [4]. If we would like to change records, we have to spend a massive premium. This approach can be used for secure transactions.

1.2 The Objective of the Work

The research work aims to provide a checkpoint approach to secure transactions using blockchain. Following are the objectives of the proposed research work.

- To study and analyze data mining techniques for fraud detection in the banking sector and their collective analysis from past experiences.
- To design a blockchain-based model using a checkpoint approach for performing secure transactions.
- To investigate and design an automatic notification to detect the unknown activity within transactions.
- To investigate/identify performance analysis based on parameters to improve the accuracy of the proposed system.

2 Literature Review

Most researchers are using blockchain technology to make a secure system, but the least amount of the literature is available to explore the blockchain-based checkpoint approach. Thus, the attempt is made to find out the related literature to explore a better security system.

To develop Industry 4.0, which is restricted by various restrictions for network performance and solid protection, Javaid and Sikdar [1] researched Industrial Internet of Things (IIoT). However, blockchain is not appropriate for addressing these with conventional proof of work (PoW) agreement in the existing construction. The researchers propose an innovative blockchain architecture that utilizes delayed proof of work (dPoW) and a control method. Contrary to conventional PoW-based mining frameworks, dPoW constructs different levels of mining complexity, enabling the infrastructure to grow effectively. In our research, the operating system specifies a stable method to produce its following hashing key in the ledger, thus raising the sophistication of the assault and improving the architectural design safety. Therefore, this property and the subsequent changes make it possible for the augmentations to ramp up and handle the growing transfer flow linked to corporate systems.

Rambola et al. [2] addressed data mining as the full range of crucial facts that change all top-ranked banking and business decisions. Individuals combine the various information from numerous repositories and save the combined information in an appropriate layout that uses a storage service so that information can be mined for this purpose. The analysis is performed more, and then, the collected data is used to assist strategic thinking throughout the financial industry or in some enterprises. Data mining tools are a significant aid to them through the financial sector to attract and manipulate potential consumers, most effective when interfering with theft, providing phase-based products, fraudulent activities, risk control, and consumer research in virtual environments. The processing of information is a crucial method for identifying crimes in banking details or crime prevention. Data mining supports the repository, improves power generation and choices, takes sensible decisions, and selects rational decisions. It retrieves a broad repository trend that aims to enhance storage consistency. Therefore, its current study contains several issues related to the reliability of account details and methods of quickly overcoming the scams caused by the money system using my data analysis methods.

Xu and Veciana [5] suggested several architectures that use redundancies, where even a task is done on several databases. This study, therefore, seeks to deal with the influence of two leading causes of heterogeneity in multiple processor groupings: working hours and processor pace. The main goal and commitment were to overhaul the entire contingency pace-up feature and offer an essential insight into where work planning and redundancies with small checkpoints can mitigate uncertainty in work responsiveness. If consistent delivery of services on distributed clusters and computer systems increases, methods like ours are probably an integral component of any reasonable alternative.

Alcaraz and Lopez [6] suggested a template checkpoint focused on a shared, trusted Internet network topology. The method can handle decentralized alert reproductions that allow enough redundant data for defect identification and access control in the test points. This means that system stability could be achieved using such copies, as fully mobilized can significantly be challenged by hackers or disturbed by the complex changes triggered by the accession of younger recruits or the departure of the entire network. Chart science and command theories were used to transform the method and achieve results via repeated tests.

The findings demonstrate which power-law representations are constructive in particular to accommodate high-precision control-based methods for detecting differences in systemic reliability. However, the study also found that perhaps the crucial importance of regulators allows level transfers to become more effective than purely physical law. As for proposed development, we hope to apply the technique to more dimensions of correspondence and signal credibility. This will allow the individual to investigate new service countries in an excellent alternative to threats linked to fake information intrusion toward better opposites. To spot strange anomalies in any portion of a system, on its whole, unique anomalies identification methods should be seen as a portion of the manufacturing method. Such jobs will include studying massive databases, such that additional facilities are welcomed to share information storage. We often hope to integrate these pieces into a reduced evidence framework to understand new output metrics better and improve the research.

Ohara captures the further focus and Al [3] checkpointing by ledger data management in the inquiry. The simulators built and monitored link behavior in preparing an empirical framework for predicting the estimated overall costs associated with stalled stones checkpointing and restoration. Although in emerging cryptographic protocols for general use, there are a few uncomfortable levels. In particular, blocked frames contribute to an unwanted blockchain asset of inspection. A far more practical simulation was built to investigate freedom to convey perplexing behavior of authors that show exactly blockchain framework parameters. To minimize the complexity of systems and provide the precision of categories' comparison with several other current approaches, Mittal et al. [7] have created awareness and training that uses the master training methodology, and the built structure is built. In addition, for every consumer to be accessible as standard or otherwise, an accessibility method is needed for accurate evaluation. This work discusses various machine learning methods that prove that automatic training effectively evaluates credit risk. Apps like gender, age, profession, etc., is used to reach every positive or negative consumer. These factors are essential in determining a consumer's reputation. Thus, the machine learning technique would be carried out to improve performance.

Throughout the K-means and K-medoids process, Aryuni et al. [8] also established group versions of consumer user profiles focused on consumer Internet banking of consumer groups in XYZ. The results were calculated as contrasted with the segmentation algorithm. The system of K-means outperforms K-medoids on the ground of intra-cluster, i.e., cluster average (AWC). K-Means do marginally better than K-medoids depending on the Davies–Bouldin index (DBI). Centered mainly on the lowest DBI amount, the correct outcome in clusters in both approaches is 2.

The framework will be used to (1) combine transaction statistics, social and economical, and commodity possession details mostly on online banking networks and (2) establish the customers' law using different classifiers. This framework is recommended for further research projects.

The efficiency of the various data mining programs suggested by Başarslan and other partners [9] was analyzed by establishing categorization technique models. Four programs have shown multiple outcomes. Even so, the decision tree has been the method that produced significant consequences for both programs.

Wong et al. [10] recognize and test the utility of multiple temporary movement (LSTM) channels using a validation strategy. The purpose is to forecast big data management efficiency (DQ) or improve the prospective capacity of DQ for the financial sector. That model of significant banks would enhance their ability to fulfill the Basel Banking Supervisory Committee's global legal standards (BCBS 239). CPG 235 is intended to help controlled organizations manage data risk in Australia. It is also structured to direct top executives, risk management, and strategic experts (respectively administrative and operational). The model may also be used to comply with other related standards, such as APRA CPG 235.

Individuals apply this approach to track danger disturbances via a world financial efficiency benchmark and calculate the effects using Gaussian and Bayesian techniques. Furthermore, we conduct concurrent training via a focus function and estimation in many artificial neural networks. That system is tested using different service methods to demonstrate machine learning technology's prediction performance and verified to verify the system's efficiency.

Kostić et al. [11] proposed an approach that should incorporate technical understanding and simulation of data science and a financial challenge. Through lowering prices in advertising agencies, the findings obtained could be helpful. Furthermore, the results provided will be used to make customer communication more effective. There will be a significant improvement in the number of favorable replies to customer encounters in many other terms. It is crucial to understand that information collected by differentiation should be used in conjunction with some other form of customer tracking or assessment.

The difference between blockchain and bitcoin was demonstrated by Tasatanattakool and Techapanupreeda [12]. They further addressed how an illegitimate (unauthorized shareholder) entity would view or apply to hospital or medical authorities for a physician's medical information while breaching the individual's confidentiality.

The research paper offers an extensive description of blockchain technologies. Zheng et al. [13], the readers, summarize blockchain but first relate a few other standard methodologies of agreement used during various blockchains. In addition, there are short descriptions of technical problems or significant upgrades. The writer even proposes any likely new guidelines. Blockchain-based technologies are currently emerging, and we expect the future to carry out thorough research into blockchains.

Ferrag et al. [14], throughout the work, examined the latest technology of current IoT blockchain rules. We presented a summary of blockchain IoT technology applications such as IoT, IoE, IoC, and edge processing. Writers became capable of identifying the hazard models which blockchain protocols regarded within IoT networks across five major categories by detailed study and analyses. Many challenge fields remain, like cumulative resilience, the complex and adjustable protection architecture, energy-efficient extraction, social networking sites and confidence management, blockchain-specific technology, distribution of cloud vehicle advertising, and skyline question treatment.

Discrete logarithms or structures are dependent on RSA. An Ethereum-based framework has been used to verify the effectiveness of the suggested software.

Kavitha and Saraswathi [4] address various methods of encryption and decryption of images. Different encryption methods are researched and analyzed to support the encryption recital. The original picture is incorporated and encrypted in all processes and then sent to the recipient. Every algorithm is unique, and the method is uniquely implemented. The new technique of encryption is changing every day. High-security encryption methods are still working out.

Tahir and Javaid [15] have proposed a decentralized way to share big data. This approach aims to build an environment where all participants will participate in the peer-to-peer sharing of data. The central part is to use blockchain technology to archive transaction records and other essential documents. The solution does not need third parties, unlike current data exchanges markets. It also provides data owners with a simple way to verify the use of data and protect the rights and privacy of data.

Sokouti et al. [16] have used Goldreich Halevi Goldwasser (GGH) algorithm in the numerical frames for the encryption of medical images, the GGH algorithm does not increase the image size, and thus, its complexity remains as straightforward as O(n2). However, the chosen ciphertext attack is one of the drawbacks of using the GGH algorithm. This deficiency in the GGH algorithm was resolved and improved in this strategy by using padding (i.e., snail tour XORing) before the GGH encoding process.

Ramani et al. [17] have proposed a secure mechanism for access to information, which can provide patients with access only by authorized entities. The author, therefore, considers that blockchain technology safeguards data on healthcare systems as a distributed approach. The general key-related operations in the proposed scheme rely on elliptic curve cryptography (ECC), offering more lightweight available essential cryptographic functions than the classical discreet logarithms or structures dependent on RSA. An Ethereum-based framework has been used to verify the effectiveness of the suggested software.

Liu et al. [18] have suggested a data-sharing network known as BPDS based on blockchain privacy. In BPDS, the initial EMRs are safely stored in the cloud, and the indexes are held in a manipulated blockchain consortium. This will dramatically reduce the risk of medical data leakage while, at the same time, ensuring that blockchain indexes cannot be unilaterally updated throughout the EMRs. According to the predefined patient access privileges, secure data sharing can be done automatically via intelligent blockchain contracts. Furthermore, the CPABE-based framework for access controls and the content signature extraction scheme ensure that data sharing maintains good privacy. Security analysis shows that BPDS is a secure and efficient way to share EMR data.

Kim et al. [19] have proposed the Distributed Machin Learning framework for just a licensed blockchain systemically build a separate, stochastic gradient descent approach and an errors-based accumulation principle as fundamental primitives to address safety, safety, and efficiency problems. The suggested mistake-based accumulation principle works to stop assaults of an opponent node which attempts to alter the exactness of DML models. The template is less complicated in terms of effort.

The novel DPS for medical information was suggested in Li et al. [20]. Methods used in this data preservation system are cryptographic algorithms such as secure hash algorithm (SHA-256) and elliptic curve cryptography (ECC). The advantage is that these unchanging cryptographic and memory management algorithms help to handle the leaked data. But there is a need to optimize the structure of data stored in the blockchain and image content recognition.

Li et al. [21] have developed a ring signature-based blockchain confidentiality system. This strategy uses the full ring signature's confidentiality to ensure data and user identification protection in blockchain technology to create a safe data storage mechanism based chiefly on ring signature, mostly on elliptical curves.

Liu et al. [22] proposed a secure management framework for the digital certificate-based access to blockchain data. The approach uses the mix of blockchain and digital certificate technology, which develops a protected data protection verification protocol in blockchains without checking the third-party participant's encrypted identification signature. There is a need to explore different theories in large-scale transactions, such as homomorphic attribute encryption and storage space balance.

3 Research Gap Identified

In our daily lives, the banking industry has tremendous importance or meaning. During purchases, clients and the banking will encounter several challenges due to scammers and offenders, although there are more significant risks of being stuck. And therefore, it is pretty crucial to monitor such scams.

The fundamental problem in the provision of banking includes the lack of proper data protection and security that is how data can be checked and transmitted.

There are already more than millions of records in most systems, so maintaining data ownership and privacy is a significant concern without tampering with data sets.

There are restrictions on remote access and review personnel.

Sometimes, it is inaccurate or not able to get any information due to unavailability of online connection.

Maintaining transaction-related data online is a big issue without proper security.

There are issues related to data interoperability, scalability, and privacy.

4 Problem Statement

For the daily lives, the financial industry has tremendous importance or meaning. Every entity gets similar offline, (1) physical and (2) virtual use of its financial system.

- There could be actual theft such as credit card stolen and financial data shared with unethical bank staff.
- Digital deception is carried out by exchanging information about the card, mainly on the Internet or mobile. Spamming and phishing can also be included.

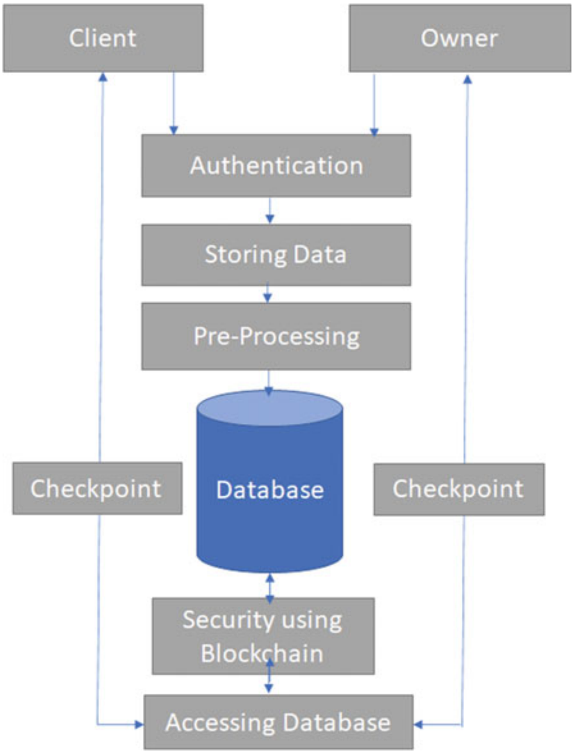
Due to scammers and terrorists, clients and lenders will face numerous challenges when doing the transfers. All the relationships with the bank's policies and the risks of being caught are far more significant. Such scams often include credit card scams, identity theft, tax fraud, and so on, which can add to banking or consumers' economic damages. And therefore, it is essential to detect such fakes.

5 Proposed Work

In different fields, blockchain innovations are more and more popular. Many clients' aspect applications now store information in ledger processing. The data is protected, traceable, and tolerant to error—many applications using blockchain-based digital data for trustworthiness in the coming years. Checkpoints have been commonly used to improve reliability. Throughout the paper, the impact of stalled blocks on dispersed controls is discussed (Fig. 2).

The research aims to develop an empirical method to forecast the overall risks and efficiency of recovering a decentralized checkpoint strategy utilizing blockchain-based processing to ensure that data sets are appropriately authenticated digitally.

Fig. 2 Proposed framework using blockchain with checkpoint



So once again, throughout this project schedule, a simulation will be developed to monitor the actions of a blockchain device and achieve specific criteria for securing the data traffic against illegal disclosure or monetary manipulation using the developed solution.

5.1 Research Methodology Tools and Techniques

The proposed work is planned to be carried out in the following manner.

1. Collection of data related to data mining techniques for fraud detection in the banking sector.
2. Comparative analysis for selection of suitable blockchain-based model.
3. Study of checkpoints and selection of optimal checkpoints.
4. Study for automated notification for detecting the activity in a transaction.
5. Model formulation and performance analysis.
6. Model testing for optimal results.
7. Interpretation and performance analysis w.r.t. accuracy and security.

6 Conclusion

The majority of the crucial data is gained through this large quantity of data mining technology and changing all high-level choice strategies in the banking and retail industry. Individuals combine the various information from numerous repositories to save the combined data in an appropriate layout using a storage service so that information can be mined for this purpose.

Data analysis was conducted more, and then, the collected data was used to assist strategic thinking in the financial industry or some enterprise. In the financial sector, data mining technology is a massive contribution to focusing and exploitation of new customers, majorly used in scam intervention.

Data mining is a crucial method to identify scam activity in banking and deter fraud from occurring by scammers in their everyday lives. Data mining supports the repository, improves power generation and choices, takes intelligent decisions, and selects rational decisions. It retrieves a broad repository template that aims to enhance repository consistency.

Throughout this research, the consequences of blocked blocks are to be discussed on decentralized controls and actively involved in building an empirical analysis to predict downtime and retrieval effects of a decentralized checkpointing strategy leveraging distributed ledgers processing to encrypt the payment.

References

1. Javaid U, Sikdar B (2020) A checkpoint enabled scalable blockchain architecture for industrial Internet of Things, 1551-3203 (c) 2020 IEEE
2. Rambola R, Varshney P, Vishwakarma P (2018) Data mining techniques for fraud detection in banking sector, 978-1-5386-6947-1/18/$31.00 ©2018 IEEE
3. Ohara M (2019) A study on checkpointing for distributed applications using blockchain-based data storage. 2473- 3105/19/$31.00 ©2019 IEEE
4. Kavitha PK, Saraswathi PV (2017) A survey on medical image encryption. In: ICASCT2501 | ICASCT | March-April-2017 [(3) 5 : 01–08]
5. Xu H, de Veciana G (2018) Online job scheduling with redundancy and opportunistic checkpointing: a speedup-function-based analysis. 1045–9219 (c) 2018 IEEE
6. Alcaraz C, Lopez J (2017) A cyber-physical systems-based checkpoint model for structural controllability. 1937–9234 © 2017 IEEE
7. Mittal A, Shrivastava A, Saxena A, Manoria M (2018) A study on credit risk assessment in banking sector using data mining techniques. 978-1-5386-5367-8/18/$31.00 ©2018 IEEE
8. Aryuni M, Madyatmadja ED, Miranda E (2018) Customer segmentation in XYZ bank using K-means and K-medoids clustering. 978-1-5386-5821-5/18/$31.00 ©2018 IEEE
9. Başarslan MS, Argun RD (2018) Classification of a bank data set on various data mining platforms. 978-1-5386-5135-3/18/$31.00 ©2018 IEEE
10. Ka Yee W, Wong RK (2020) Big data quality prediction on banking applications: extended abstract. 978-1-7281-8206-3/20/$31.00 ©2020 IEEE
11. Kostić SM, Đuričić M, Simić MI (2018) Data mining and modeling use case in banking industry. 978-1-5386-7171-9/18/$31.00 ©2018 IEEE
12. Tasatanattakool P, Techapanupreeda C (2018) Blockchain: challenges and applications. 978-1-5386-2290-2/18/$31.00 ©2018 IEEE
13. Zheng Z, Xie S, Dai H, Chen X, Wang H (2017) An overview of blockchain technology: architecture, consensus, and future trends. 978-1-5386-1996-4/17 $31.00 © 2017 IEEE DOI https://doi.org/10.1109/BigDataCongress.2017.85
14. Ferrag MA, Derdour M, Mukherjee M, Derhab A (2018) Blockchain technologies for the Internet of Things: research issues and challenges. 2327-4662 (c) 2018 IEEE
15. Tahir J, Javaid N (2017) Bootstrapping a blockchain-based ecosystem for big data exchange. 978-1-5386-1996-4/17 $31.00 © 2017 IEEE DOI https://doi.org/10.1109/BigDataCongress.2017.67
16. Sokouti M, Zakerolhosseini A, Sokouti B (2016) Medical image encryption: an application for improved padding based GGH encryption algorithm. Open Med Inform J 10:11–22
17. Ramani V, Kumar T, Braeken A, Liyanage M, Ylianttila M (2018) Secure and efficient data accessibility in blockchain-based healthcare systems. https://doi.org/10.1109/GLOCOM.2018.8647221
18. Liu J, Li X, Yey L, Zhangy H, Duz X, Guizanix M (2018) BPDS: a blockchain-based privacy-preserving data sharing for electronic medical records. 978-1-5386-4727-1/18/$31.00 ©2018 IEEE
19. Kim H, Kim S-H, Hwang JY, Seo C (2019) Efficient privacy-preserving machine learning for blockchain network. IEEE Access 7:136481–136495
20. Li H, Zhu L, Shen M, Gao F, Tao X, Liu S (2018) Blockchain-based data preservation system for medical data. https://doi.org/10.1007/s10916-018-0997-3
21. Li X, Mei Y, Gong J, Xiang F, Sun Z (2020) A blockchain privacy protection scheme based on ring signature. IEEE Access 8:76765–76772
22. Liu B, Xiao L, Long J, Tang M, Hosam O (2020) Secure digital certificate-based data access control scheme in blockchain. IEEE Access 8:91751–91760

Chapter 13
Analysis on Interaction of Machine Learning with BlockChain

Gayatri Gattani and Shamla Mantri

1 Introduction

The various large and important industries of modern society such as healthcare, agriculture, IT, business, research and development, security depends on the system that predicts outcomes, analyzes the growth, degradation, behavioral patterns, and automates the workflow with ease and efficiency. The shaping of such a system is possible with the use of different algorithms in machine learning. However, machine learning is a system that is dependent upon data for its learning. It extracts the features from data and trains the model itself. Big data has made it feasible to build effective machine learning models that detect, predict, classify and obtain rigorous knowledge about the various circumstances in different fields with high accuracy [1]. Along with the volume of data, consistency and integrity of data are equally pertinent to get true and efficacious results as well as predictions.

The data used by the average person on the Internet is estimated nearly to 2.5 quintillion bytes per day according to 2020 reports. This kind of massive amount of data needs to be stored securely such that no invalid third person or authorization can corrupt or poison this data. Here, we consider the concept of blockchain technology that has a decentralized database system. Blockchain is a ledger that connects peer-to-peer networks without the need for a third-party intermediary where every user on the network is provided with a duplicated copy of the ledger. It is a decentralized system with the unique features of transparency and immutability [2].

Thus, the use of blockchain in machine learning can provide a better and more secured form of datasets that will build the systems more reliable and worthy. Blockchain stores data in the structured blocks that are chained together. For the entry of new data, a new block is created. Every block has data, a unique hash value, and the hash value of the previous block. This is how the chain between blocks is

G. Gattani (✉) · S. Mantri
School of Computer Engineering and Technology, MIT World Peace University, Pune, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_13

formed. The use of ML in administrating the chains can upgrade the security of the blockchain network [3]. The interaction of machine learning technology with blockchain can revolutionize emerging industries.

2 Literature Survey

2.1 Data

Data is the base of new emerging methodologies to ameliorate the efficiency of the systems, make predictions based on changes and outcomes, and analyze the growth and degradation pattern of the sector in a company or industry. Without data, there is no purpose left for the models to be built. The substantial amount of precised data results into proper ratio of training and testing set that ensures the accuracy results and predictions out of models.

2.2 Machine Learning

ML is a part of artificial intelligence (AI) and software engineering which centers around the utilization of information and calculations to copy the way that people learn, slowly working on its precision. Classification and predictions are made on the basis of calculations that are prepared to make groupings or expectations, revealing key experiences inside information mining projects. As of now, the field of Ml is coordinated around three essential research foci as mentioned in the Table 1.

Machine learning is completely based on the datasets. It just needs the training datasets and the algorithm used itself learns the classification rules from the set of training samples from datasets. The best possible outcome for any algorithm can be checked using various evaluation techniques in machine learning. This evaluation technique also helps in recognizing the overfitting and underfitting of the model. Cross-validation score, F1 score, confusion matrix, precision, recall, accuracy, regression metrics, mean squared error can be used for evaluating the model.

Table 1 Description of primary research focus of machine learning [4]

Focus	Description
Task-oriented studies	The turn of events and examination of learning frameworks to further develop execution in a foreordained arrangement of assignments (otherwise called the "designing methodology")
Cognitive simulation	The examination and virtual experience of human learning measures
Theoretical analysis	The theoretical investigation of the space of conceivable learning techniques and calculations free of use area

The three major metrics to weigh up a classification model are accuracy, precision, and recall [5].

2.3 Blockchain Technology

A blockchain is characterized as a reliable and secure decentralized and circulated network that gives collaborations among members, for example, networks that comprise of people, organizations, or governments that have a particular or shared objective. Every member in the blockchain has a common record with others to ensure permanence and consistency for each exchange, with every exchange checked for legitimacy by an agreement of a larger part of hubs [6]. Thus, blockchain generated data is secured and valuable as well, meaning it is organized, plentiful and complete, making it an ideal hotspot for additional analysis. Blockchain has brought an entirely different method of overseeing and working with information—as of now not in a focal point of view where all information ought to be united, however in a decentralized way where information might be broke down directly off the edges of individual gadgets [7].

2.4 Interaction of Blockchain Technology and Machine Learning

Blockchain empowers secure capacity and sharing of information, and ML can examine and create experiences from such information to produce esteem. As such, we can say that ML can be utilized to recognize the check the accuracy of information, and once approved, it very well may be sent to the blockchain organization to make the data invariable. ML is known to deal with enormous information; along these lines, it offers a chance to assemble better AI models by utilizing the decentralized idea of blockchains. Keeping a store network, in pragmatic, is a perpetual assignment for all reach endeavors, and the interconnectivity of numerous components in the store network continually turns out to be more wasteful when a business develops. Blockchain can be carried out to many difficulties of the supply chain industry. Blockchain works on the productivity and straightforwardness of the inventory network, making it conceivable to follow everything from warehousing to conveyance. Besides, it oversees monetary exchanges, too. The analysis using machine learning models can be made on different attributes of this supply chain network that ease the workflow of complete process and saves time [8].

A clinical analyst who needs to give patient-explicit therapy can prepare a prescient model of infection by working together with secure clinical networks in a blockchain network with no extra course of haggling with one another for an information base [9]. To run a learning model without information centralization, a distributed machine learning (DML) model for blockchain organizations ought to be assembled dependent on numerous elements, i.e., a figuring hub and different specialists for equal handling. This implies that the figuring hub processes worldwide load by gathering just taking in outcomes from every specialist as a member in each round. A DML can give viable methods of utilizing information innately dispersed across numerous spaces without a focal information worker [10, 11].

Blockchain technology has gone a long ways past bitcoins. Medical services is one of its application regions. The proposed framework depended on bitcoins approach was addressing data client's need and ensuring patient's protection [12]. The significant component of blockchain innovation is the improvement of safety and protection without the contribution of an outsider authenticator. The blockchain network is for the most part to be considered as secure and versatile yet their security level is straightforwardly corresponding to the measure of hash processing power that upholds the blockchain. Blockchain can be attacked in various different ways [13]. There are versatile ML algorithms like support vector machines (SVM), clustering, and deep learning (DL) algorithms such as convolutional neural network (CNN) that are used to analyze the attacks on blockchain-based networks [14].

The learning abilities of ML can be applied to blockchains-based applications to make them more astute. By utilizing ML security of the conveyed record might be improved. ML may additionally be utilized to upgrade the time taken to arrive at agreement by building better information sharing courses. Further, it gives freedom to assemble better models by exploiting the decentralized engineering of blockchain technology [14]. With the current world being reliant upon information driven investigation requiring exact ML calculations, it becomes incredibly important to guarantee guard from all potential assaults. There is a critical need to assemble a model that is adequately vigorous to battle all such assaults on datasets and ML calculations [15].

3 Conclusion

The stage, where data is used everywhere to analyze growths and statistics, integration of two massive technologies that is machine learning and blockchain technology is considered revolutionary for all the emerging industries of IT, health, business and research. The results obtained by the ML models using blockchain-based data ensures accuracy since the blockchain technology has data in its secured and organized form. Blockchain technology has gone a long ways past bitcoins. Blockchain can be carried out to many difficulties of the supply chain industry. There are versatile ML algorithms that are used to analyze the attacks on blockchain-based networks. There are various advancements being done in the combination blockchain and machine learning.

4 Future Work

Bitcoin is not the only application of blockchain technology. Number of opportunities in the field of blockchain can be studied to enhance the growth of this technology. The rise of data and machines is increasing day-by-day. Practical implications to test and compare the accuracy of ML models using the blockchain driven database and a normal database can be done. Find the sector, where the bulk data can be stored in the blockchain in organized form and the process of automation is eased with the help of using ML. Study on various attacks on data that can be resolved using blockchain so as to have noise free data as input in ML model can be studies further.

References

1. Acheampong F (2018) Big data, machine learning and the blockchain technology: an overview. Int J Comput Appl 180:1–4. https://doi.org/10.5120/ijca2018916674
2. Leible S, Schlager S, Schubotz M, Gipp B (2019) A review on blockchain technology and blockchain projects fostering open science. Front Blockchain 2
3. What happens when you combine blockchain and machine learning. https://medium.com/@Intersog/what-happens-when-you-combine-blockchain-and-machine-learning-2afafc9654d2
4. Carbonell JG, Michalski RS, Mitchell TM (1983) An overview of machine learning. Mach Learn 3–23. https://doi.org/10.1016/b978-0-08-051054-5.50005-4
5. Evaluating a machine learning model. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/
6. Kim H, Kim S, Hwang JY, Seo C (2019) Efficient privacy-preserving machine learning for blockchain network. IEEE Access 7:136481–136495. https://doi.org/10.1109/ACCESS.2019.2940052
7. How blockchain will disrupt data science: 5 blockchain use cases in big data. https://towardsdatascience.com/how-blockchain-will-disrupt-data-science-5-blockchain-use-cases-in-big-data-e2e254e3e0ab
8. Is blockchain with Ml a great combination? https://www.blockchain-council.org/blockchain/is-blockchain-with-ml-a-great-combination/
9. Hynes N, Dao D, Yan D, Cheng R, Song D (2018) A demonstration of sterling: a privacy-preserving data marketplace. VLDB 2018, pp 2086–2089
10. Hamm J, Cao Y, Belkin M (2016) Learning privately from multiparty data. In: International conference on machine learning. New York, NY, USA, pp555–563
11. Bellet A, Guerraoui R, Taziki M, Tommasi M (2018) Personalized and private peer-to-peer machine learning. In: Proceeding of the 21st international conference on artificial intelligence and statistics. Playa Blance, Lanzaroto, Canary Islands, pp 1–20
12. Tanwar S, Bhatia Q, Patel P, Kumari A, Singh PK, Hong W (2020) Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward. IEEE Access 8:474–488. https://doi.org/10.1109/ACCESS.2019.2961372
13. Gadekallu TR, Manoj MK, Krishnan S, Neeraj Kumar S (2011) Thapar Institute of Engineering, Punjab Saqib Hakak University of New Brunswick, Canada Sweta Bhattacharya. Blockchain based attack detection on machine learning algorithms for IoT based E-health applications, arxiv (2011), 2011.01457v1
14. Tasatanattakool P, Techapanupreeda C (2018) Blockchain: challenges and applications. Int Conf Inf Netw (ICOIN) 2018:473–475. https://doi.org/10.1109/ICOIN.2018.8343163
15. Aggarwal S, Kumar N, Raj P (2021) Attacks on blockchain. Adv Comput 121:399–410

Chapter 14
Frequency Response Masking Based Reconfigurable Filter Bank for Hearing Aids

Anjali Shrivastav and Mahesh Kolte

1 Introduction

Hearing loss disrupts interpersonal communication and disturbs human relations. WHO reports—“More than 5% of the total population of world has hearing loss. This count has raised from 360 million which existed before five years. It is further projected to rise to over 900 million people by year 2050 [1].

Hearing impaired people who range between mild and severe hearing loss can take advantage of hearing aids. Untreated hearing impairment may also lead to symptoms similar to Alzheimer’s disease and patient may develop Dementia. They may also experience earlier cognitive abilities impairment and reduced thinking abilities compared to adults with normal hearing. Statistics shows even in developed countries hardly 16% adults and 30% age old hearing impaireds (HI) actual use aids. The tremendous rise in count of hard of hearing people and significant health problems suggest need for immediately addressing need to further improve available hearing aids to boost the count of actual hearing aid users to avoid further related health issues of affected community of people with hearing misfortune.

The most commonly occuring Sensorineural Hearing Loss (SNHL) is because the cilia hairs within cochlea fail to differentiate between different frequency cues unlike people with normal hearing. They may also suffer from “Loudness recruitment” problem causing irregular loudness impression. The type and degree of hearing loss along with their impact on hearing impeded individual is in Fig. 1.

The hearing profile through audiometry test is received in form of an audiogram which displays hearing threshold of individual at octave frequencies from 250 to 8 kHz. The audiogram projects hearing profile in terms of degree of hearing loss

A. Shrivastav (✉) · M. Kolte
E and Tc Department, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune, India
e-mail: anjali.shrivastav@pccoepune.org

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_14

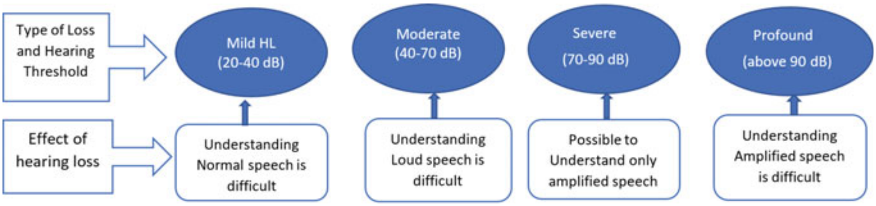

Type of Loss and Hearing Threshold
Mild HL (20-40 dB)
Moderate (40-70 dB)
Severe (70-90 dB)
Profound (above 90 dB)
Effect of hearing loss
Understanding Normal speech is difficult
Understanding Loud speech is difficult
Possible to Understand only amplified speech
Understanding Amplified speech is difficult

Fig. 1 Types of hearing loss with their hearing threshold and impact

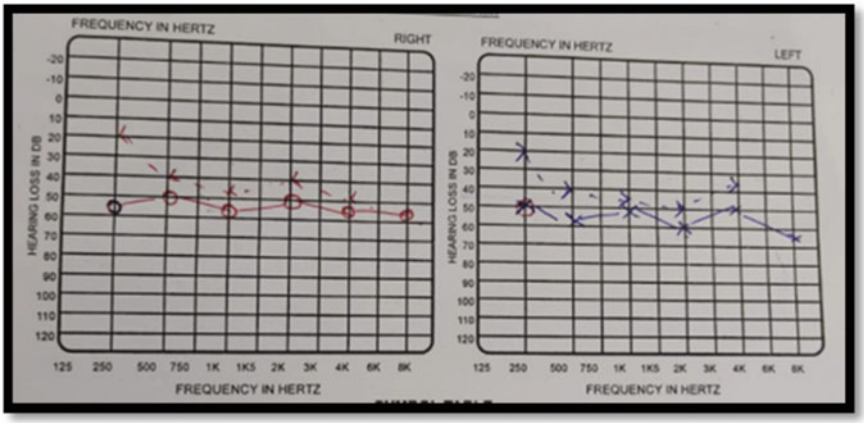


Fig. 2 Sample audiogram of hearing ımpaired with moderate SNHL hearing loss

ranging from Mild to Profound as per hearing threshold at octave frequencies as in Fig. 2.

Different types of audiometry problems can be related to understanding capability reduction if loss is less than 45 dB and above that may lead to problem in differentiating between cues [2]. Other problems pertaining to hearing loss may include reduced dynamic range, elevated hearing threshold leading to loss of sensitivity to weak sounds, diminished frequency resolution, down fall in temporal resolution and spatial cues causing bad impact on speech intelligibility of individual.

Digital hearing aid comprises of an anti-aliasing filter which post conversion into digital form through A/D converter is fed to filter bank which then decomposes the input into varying sub-bands and further provides gain selectively to each band for improved audiogram matching. The output of filter bank is once again converter back to analog form using D/A converter and is applied to reconstruction filter for ensuring close to perfect reconstruction. The Digital Hearing Aid Structure presented in Fig. 3.

Hearing aids improvises the hearing quality of patients with hearing loss ranging from mild to severe. Through selective amplification of sound, it can make speech more intelligible. Figure 4 depicts the research and inventions done so far for

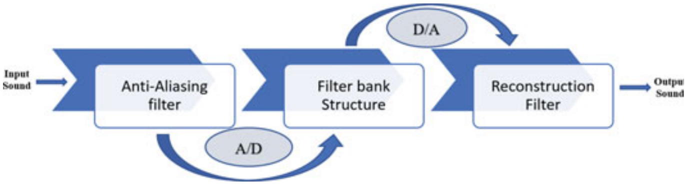


Fig. 3 Overview of digital hearing aid structure

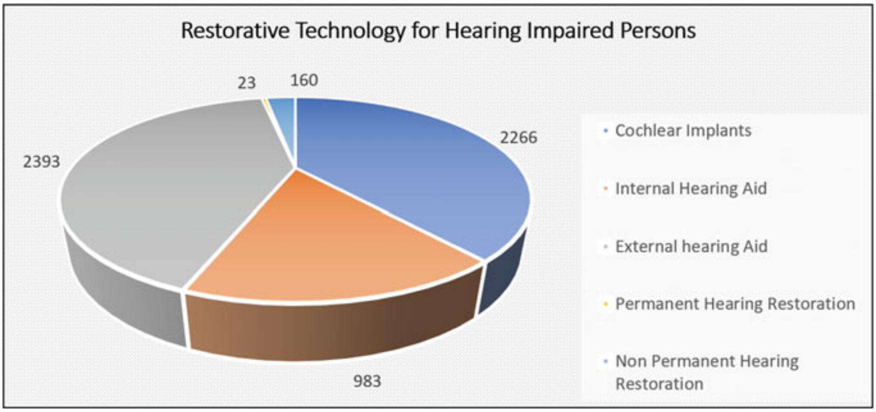


Fig. 4 Assistive devices ınventions for hearing ımpaired [3]

benefitting HI as per the WIPO statistics [1].

In spite of the myriad of research and inventions available for hearing restoration of impaired, exhaustive literature survey brings to the forefront a variety of challenges in hearing aid design including the need for narrow transition bandwidth with minimum hardware complexity and least possible overlapping in between bands for better frequency selectivity, need for improved speech quality, reduced size for improvising aesthetics, lesser power dissipation for easy portability purpose, noise reduction and higher SNR in noisy ambience and many more.

2 Reconfigurable Filter Bank for Hearing Aid Application

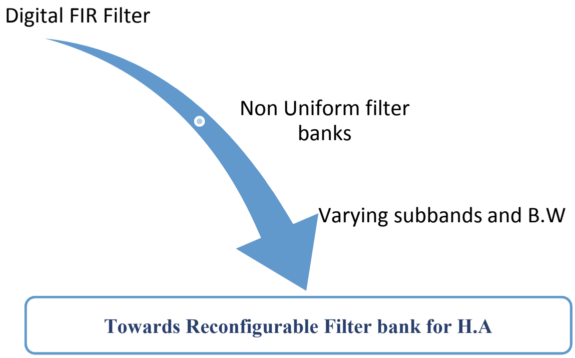
The filter bank comprises of Analysis, processing and synthesis blocks. Optimum performance of hearing aid is largely dependent on suitable filter bank used. In comparison to analog, digital filters are more apt owing to their flexibility, parallelism feature, programmability, accuracy [4]. Finite Impulse Response (FIR) is more stable and has no phase distortion linear phase response. The non-uniform filter is more suitable for the hearing resolution properties of human. Filter banks with fixed number and bandwidth of sub bands reduce the flexibility of matching with audiogram restricting auditory compensation improvement [5]. Thus, variable Bandwidth filter gives better matching and compensation compared to fixed, non-uniform filter bank [1].

Thus FIR filter banks (FB) which is Non-Uniform in nature with changing subband count and varying BW is good option for Hearing aids as shown in Fig. 5 and researchers can work towards designing filter bank structure with improvement in performance parameters including improved matching error, better flexibility, reduced power consumption and dissipation and lowering of Hardware and computational complexity with major focus on reduction in delay, size and improved sound quality and intelligibility.

Reconfigurability in hearing aid is the need of the hour as majority of hearing-impeded individuals looks for customized and tailormade hearing aids flexible enough to take into account the variable hearing needs which is possible through the design of reconfigurable filter bank (RFB). The Generalized structure of reconfigurable FB is depicted in Fig. 6.

So, need is to use same hearing aid device which will satisfy varying needs of H.I. The filter bank should be able to select suitable Frequency range where patient has low hearing sensitivity. This reconfigurability feature can be very important step towards tailor-made hearing aids design.

Fig. 5 Filter bank types and preferences for hearing aid application



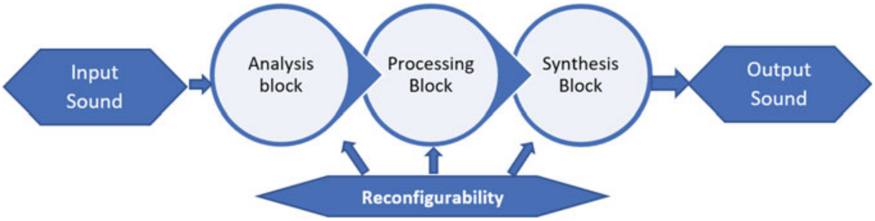


Fig. 6 Generalized structure of RFB

Some of the major design challenges in reconfigurable filter bank design includes the need for adjustable and multiple band spectral sub band decomposition schemes with elevated tuning flexibility and narrow buts harp transition bandwidth while maintaining equilibrium between filter performance and design complexity. Reconfigurable filter bank design allows reconfiguring of various filter bank parameters including dynamically varying of sub band bandwidths [6], changing the filter coefficients, altering sub-band schemes, selecting different bandwidth sets and spectral shifting of selected sub-bands [7], varying cut off frequencies of low pass, band pass and high pass filter [4], adjusting the number of sub bands along with their location [8], flexibly controlling the bandwidth and central frequencies of filter bank [9] and changing TF of filter bank by changing value of control signals and thus giving various sound decomposition schemes [10].

3 Proposed Frequency Response Masking (FRM) Based Reconfigurable Filter Design

The FRM approach provides sharp transition bands with reduced coefficients count, better computation speed and reduced power consumption. FRM strategy reduces overall complexity by decrease in the number of multipliers and adders [11]. A low-power, small size and computationally efficient nonuniformly spaced 16 subband filter bank for hearing aid applications is presented by Wei et. al. [2]. The use of band-edge shaping filters with FRM technique reduces the computational complexity of FB with improved group delay and power consumption. To reduce computational costing of FIR, Wei et. al. [12] designed FB with 2 prototype half-band filters and the sub-bands with midpoint symmetry. Using FRM methods, author can achieve audiogram matching within 5dB. Zhang et. al. [13] used FRM approach for reducing computational complexity of proposed unified complex modulated Filter bank with narrow transition bandwidth to achieve varying even or odd-stacked, maximally or non-maximally decimated design. Also, it can be directly applied to high sampling rate systems. Sebastian et. al. [14] proposed low complex non-uniformly spaced FIR filter bank for digital hearing aid application where with FRM technique and half-band filter, a drastic reduction in the number of multipliers and adders in linear phase FIR filter was achieved. Here FRM technique is achieved by cascading different combinations of prototype filter and its interpolated filters to produce sub-bands.

The system model for proposed FRM filter is shown in Fig. 7.

Group delay of the filter is given by Eq. 1

$$\text{G.D} = \left(\frac{N-1}{2}\right)m + \max\left\{\left(\frac{N_a-1}{2}\right), \left(\frac{N_C-1}{2}\right)\right\} \tag{1}$$

The Frequency response masking filter output is given by Eq. 2.

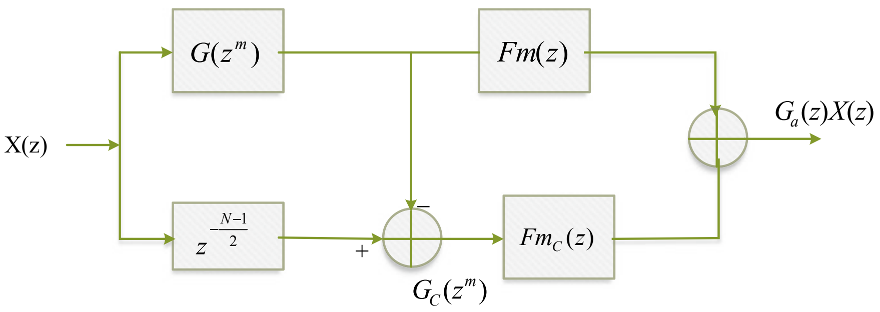


Fig. 7 System model for FRM filter

$$\boldsymbol{X(\mathbf{Z}) = G(\mathbf{Z}^m)\mathbf{Fm}(\mathbf{Z}) + G_c(\mathbf{Z}^m)\mathbf{Fm}_c(\mathbf{Z})} \tag{2}$$

The proposed sub band distribution scheme for proposed reconfigurable FRM filter bank provides flexibility to choose from a set of 3 possible schemes viz. Scheme 1, 2 and 3. The arrangement made for scheme and sub-band selection is as shown in Fig. 8.

The proposed reconfigurable FRM model is further optimized for improving performance of hearing aid in terms of better audiogram matching and reduced group delay. In the proposed work, the metaheuristic optimization algorithm improves filter bank performance. Here masking stage function is used for distributing the complete frequency range into equal bandwidth sub bands. The masking filters are represented as $\mathrm{Fm}(z)$ and $\mathrm{Fm}_C(z)$ are masking filters whereas $G_a(z)$ and $G_a(z^m)$ are the periodic model filters. In proposed model, delay components are replaced by optimization algorithm fitness function Y as shown in Fig. 9.

The T.F of proposed optimized reconfigurable FRM filter is given by Eq. 3

$$G_a(z) = G(z^m)\mathrm{Fm}(z) + \left(z^{-\frac{N-1}{2}Y} - G(z^m)\right)\mathrm{Fm}_C(z) \tag{3}$$

where N represents length of impulse response

The delay of Proposed optimized reconfigurable FRM filter is given by Eq. 4

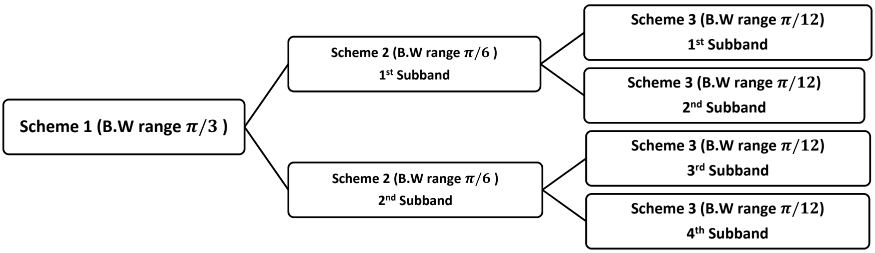

Scheme 1 (B.W range $\pi/3$)
Scheme 2 (B.W range $\pi/6$) 1st Subband
Scheme 2 (B.W range $\pi/6$) 2nd Subband
Scheme 3 (B.W range $\pi/12$) 1st Subband
Scheme 3 (B.W range $\pi/12$) 2nd Subband
Scheme 3 (B.W range $\pi/12$) 3rd Subband
Scheme 3 (B.W range $\pi/12$) 4th Subband

Fig. 8 Arrangement made for scheme and sub band selection in proposed masking filter

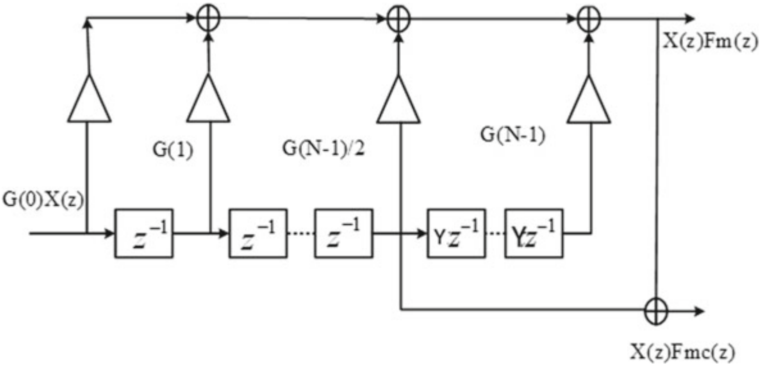


Fig. 9 Proposed reconfigurable FRM structure

$$G.D = \frac{(N-1)m}{2} + d$$
$$\text{where}\, d = \max((N_a - 1)/2, (N_C - 1)/2). \tag{4}$$

The fitness function Y is further dependent on $\mathrm{ME}_{\mathrm{max}}$ and $\mathrm{GD}_{\mathrm{max}}$ where $\mathrm{ME}_{\mathrm{max}}$ and $\mathrm{GD}_{\mathrm{max}}$ denotes optimizing factors for matching error and group delay respectively.

The proposed filter bank model is reconfigurable structure that improved the quantity of sub-bands.

The design specifications for the proposed filter design includes decomposition of the signal into 8 sub-bands. The transition bandwidth of the filter bank is considered as 0.05 and the developed filter bank has 6.4 kHz passband frequency, 16 kHz sampling frequency, 0.0001 dB maximum amount of passband ripple, and 120 dB minimum quantity of stopband reduction.

4 Results and Discussion

This section presents the single stage and two stage FRM filter designs and the frequency responses of the designed filterbank. The designed filterbank performance is compared with existing filter banks and the comparison results are presented.

The Simulink design for single stage and two-stage FRM filter design is as shown in Figs. 10 and 11.

The Frequency responses of the periodic model filter and masking filter is shown in Figs 12 and 13.

The matching error and delay of proposed reconfigurable FRM model filter is compared with the existing FRM filter designed using Signed Approximate multiplier method [15] for various ranges of hearing loss including Mild, Moderate, Mild to moderate. Severe and Profound and the results are tabulated as shown in Table 1.

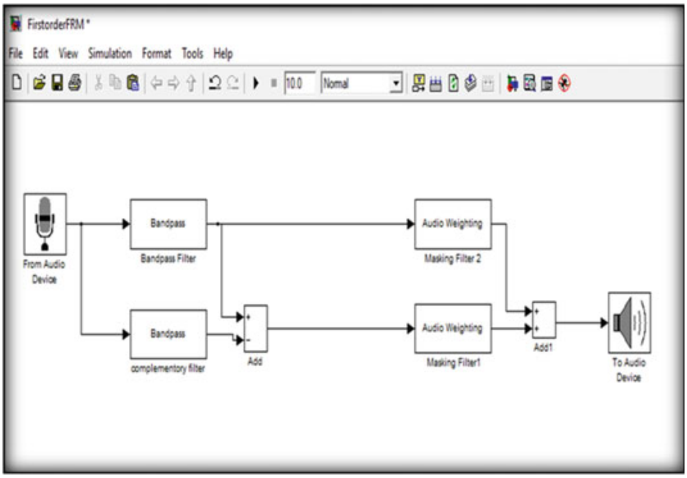


Fig. 10 Simulink model for single stage FRM filter

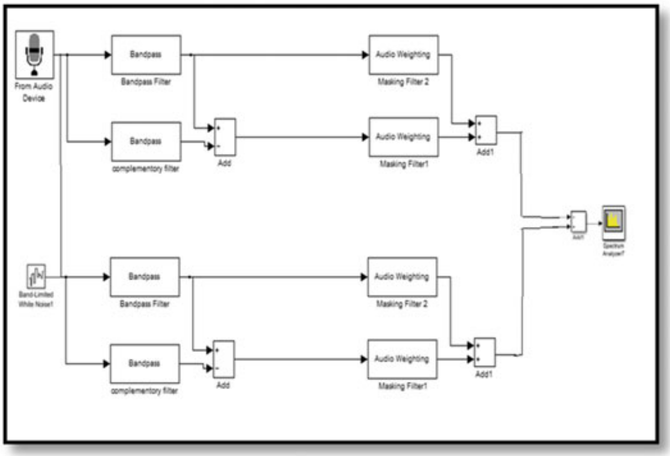


Fig. 11 Simulink model for two stage FRM filter

5 Conclusion

This paper presented the Frequency Response Masking based reconfigurable filter bank structure which provides 3 sub band distribution schemes for improved flexibility and better auditory compensation for varying needs of hearing impaired. The proposed filter using metaheuristic optimization algorithm drastically improved the matching error by almost 60% and reduced delay significantly in between 20–45% as compared to existing FRM filter used in biomedical application. Thus, the enhanced

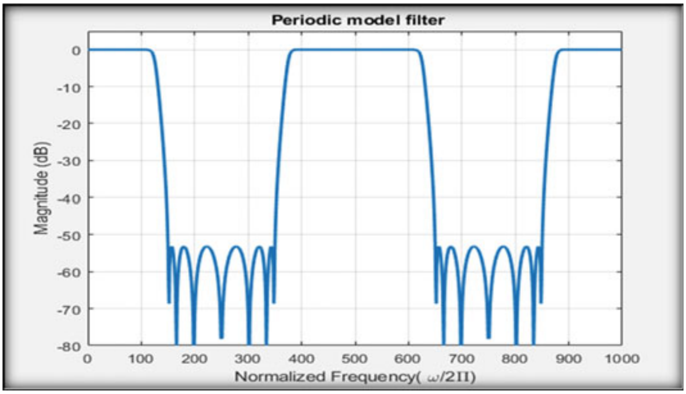


Fig. 12 Frequency response for periodic model filter

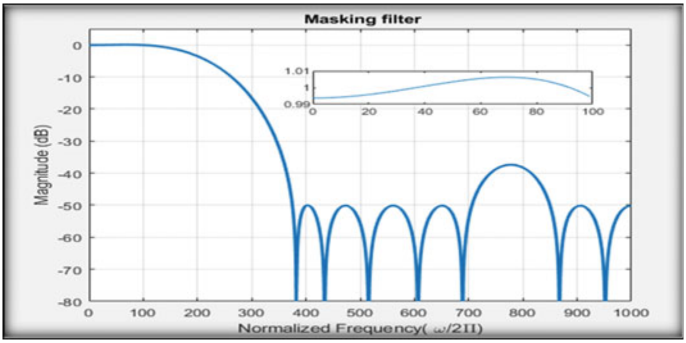


Fig. 13 Frequency response for masking filter

reconfigurable filter bank for hearing aid proposed in the paper is a stepping stone towards bridging the enormous gap between the needy and hearing aid users at actual in society.

Table 1 Performance comparison of proposed FRM filter with existing FRM filter

Hearing loss (HL) types	Matching error (in dB)			Delay (in ms)		
	Signed approximate multiplier based FRM [15]	Proposed reconfigurable FRM	% Improvement	Signed approximate multiplier based FRM [15]	Proposed reconfigurable FRM	% Improvement
Mild HL (high frequency)	3	1.2	60	4.56	2.5	45.2
Moderate HL (high frequency)	3.2	1.23	61.6	4.8	3.75	28.9
Mild to moderate HL	4.52	1.1	75.6	5.34	4.8	10.2
Severe HL	3.8	1.5	60.5	5.24	3.79	27.7
profound HL	3.5	1.45	58.6	5.2	4.68	10

Acknowledgements This work is funded by Savitribai Phule Pune University under IQAC ASPIRE Research Mentorship Grant scheme

References

1. Shrivastav A, Kolte M (2020) Reconfigurable filter bank design techniques for hearing aid performance ımprovement. Int J Recent Technol Eng (IJRTE). https://doi.org/10.35940/ijrte. Published By: Blue Eyes Intelligence Engineering & Sciences Publication
2. Wei Y, Ma T, Ho BK, Lian Y (2019) The design of low-power 16-band nonuniform filter bank for hearing aids. IEEE Trans Biomed Cir Syst 13(1):112–123
3. Solomon N, Bhandari P (2015) Patent landscape report on assistive devices and technologies for visually and hearing impaired persons. WIPO
4. Dhabu S, Vinod AP (2015) Design and FPGA ımplementation of variable cutoff frequency filter based on continuously variable fractional delay structure and ınterpolation technique. Int J Adv Telecommun Electrotech Sig Syst 4(3)
5. Nalbalwar SL, Sheth S (2015) Design of digital FIR non-uniform reconfigurable filter bank for hearing ımpairments. Int J Ind Electron Electric Eng 7:82–85
6. Haridas N, Elias E (2016) Efficient variable bandwidth filters for digital hearing aid using Farrow structure. J Adv Res 7(2):255–262
7. Haridas N, Elias E (2016) Design of reconfigurable low complexity digital hearing aid using Farrow structure-based variable bandwidth filters. J Appl Res Technol 14(2):154–165
8. Wei Y, Wang Y (2015) Design of low complexity adjustable filter bank for personalized hearing aid solutions. IEEE Trans Audio Speech Lang Process 23(5):923–931
9. Ma T, Shen C, Wei Y (2019) Adjustable filter bank design for hearing AIDS system. In: Proceedings—IEEE ınternational symposium on circuits and systems 2019-May(61671277)1–5
10. Jayeshma A, Sunderlin Shibu D (2014) An ımproved fractional fourier transform based reconfigurable filter bank for hearing aid. Int J Eng Trends Technol 10(6):276–279
11. Rajakumari R, Kalpanadevi P (2016) A reconfigurable RLS filter for hearing guide systems. IJISET 3(4):289–293
12. Wei Y, Lian Y (2004) A computationally efficient non-uniform digital FIR filter bank for hearing aid. IEEE Int Workshop on Biomed Cir Syst. S1/3/INV-S1/17, doi: https://doi.org/10.1109/BIOCAS.2004.1454116
13. Zhang W, Du Q, Ji Q, Chen T (2017) Unified FRMbased complex modulated filter bank structure with low complexity. Electron Lett 54(1):0–1
14. Sebastian A, James TG (2015) Digital filter bank for hearing aid application using FRM technique. In:2015 IEEE ınternational conference on signal processing, ınformatics, communication and energy systems, SPICES 2015, pp 4–8
15. Ramya R, Moorthi S (2019) Frequency response masking based FIR filter using approximate multiplier for bio-medical applications. Sādhanā 44(11):1–10

Chapter 15
Stereo Audio Amplifier Board Using TDA2822

Mehaboob Mujawar, T. Gunasekaran, and Jonathan M. Cansino

1 Introduction

Audio signals having low power can be amplified to a level that is suitable for driving a loudspeaker or any other output device using an electronic circuit called audio amplifier. They are used in various wireless communications and different broadcasting platforms. They are also used in different audio equipments of various kinds. It amplifies the sound wave depending upon the requirement of the user. An electronic device that increases the power, current, or voltage of a signal is called an amplifier. It amplifies any audio signal having low power to a level desired to drive a speaker. Some of the applications of these amplifiers include wireless communications and broadcasting, and in making audio equipment like hearing aids [1]. They can be of two types, weak signal amplifiers or power amplifiers. In wireless receivers, we primarily use weak signal amplifiers. Preferably, exceedingly small input signals are used as inputs to these amplifiers. The output signals generated by these amplifiers have minimal internal noise, and the signal voltage is increased by a large factor. Power amplifiers are used in broadcast transmitters or wireless transmitters. There are two important considerations need to be taken care of in case of power amplification such as power output and the efficiency. Gain, noise, and distortion are key design parameters for any audio amplifiers, mainly, because these are interdependent. An increase in gain only leads to an increase in undesirable noise and distortion in the output signal. There are two output channels in a stereo amplifier which are used to receive stereo output signal that comes from some device and gets sent to the speakers which are connected to the amplifiers. Having said that

M. Mujawar (✉)
Goa College of Engineering, Goa, India

T. Gunasekaran · J. M. Cansino
University of Technology and Applied Sciences, Muscat—Sultanate of oman, Oman
e-mail: Jonathan.cansino@hct.edu.om

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_15

the audio amplifier can be operated using TDA2822 IC which is demonstrated in this project. In this project, we will build an audio amplifier using TDA2822 IC in the stereo mode. If we have only a single audio signal which needs to be amplified, we can use the TDA2822 IC in bridge mode to combine the amplified signals from both channels. To build high-power audio amplifiers, this IC is popularly used due to its dual-channel characteristics [2]. The dual-channel audio amplification finds its integrated use in building high-power-operated audio amplifiers. This circuit is capable of taking two inputs and providing output for two speakers simultaneously. It is also commonly used as a preamplifier in stereo high-power amplifier circuits. It can deliver up to 250 milli watts output power. It consists of two inputs and two outputs. The amplifying circuit within the IC promises noise-free operation. The decoupling capacitors are used to directly couple the outputs to the speakers. Our circuit consists of one TDA2822 amplifier IC which will be able to drive two speakers with two potentiometers for volume control [3]. This circuit is capable of connecting any input having a 3.5 mm audio jack which is the industry standard audio plugs for most of the smartphones, tablets, or other handheld devices. These devices can be directly connected to the circuit through the 3.5 mm audio jack to provide required audio input for the amplifier board directly. The power supply should be with 3–15 V range because of which it is commonly used in battery-powered applications. The jumpers are used to switch between dual-channel or single-channel operation of the circuit depending upon the requirement of the user. Two electrolytic capacitors are used to filter any DC components to prevent noise and distortion in the audio signal. The two non-inverting inputs of the IC are connected to the audio signal, and the two inverting inputs of the IC are connected to the ground through filter capacitors [4]. A 4 Ω or an 8 Ω speaker can be used for the output according to the demands of the user. A speaker with lower resistance will produce an output signal having higher power and vice versa. The final goal or aim of this audio amplifier will be to reproduce the audio signal taken from the input and change the power of the output signal at the speaker along with volume control features with minimum noise or distortion and maximum efficiency possible. We can see that the circuit has minimum number of components, and it is easy to implement.

2 Literature Survey

The goal of our project is to build an audio amplifier circuit to support mobile phone devices and small speakers or headphones using minimum components and make it compact in terms of size. From paper [5], we have seen that they have used LM386 and LM3886 IC to build an audio amplifier circuit. Their circuits also need additional capacitors and transistors for its operation. But, since we need to design a circuit with less components, we used the TDA2822 IC which does not require other additional components to support its working. In paper [5], we can see that they have compared different ICs from the LM series such as LM386, LM3886, LM358, and LM358N each having its advantages and drawbacks. Again, in paper

[6], we see how LM1875 IC is used to build a similar audio amplifier circuit, but the schematic is overly complicated having a lot of transistors. One major drawback of the LM series ICs is that these circuits are prone to noise and interference and require adjusting the components to reduce this noise in the output signal. In our circuit, we see that the TDA2822 IC tackles noise and distortion quite well. The gain of LM386 circuit however is 20 which can be increased up to 200 or adjusted to a desired level using an additional capacitor. The TDA2822 can provide a gain of only up to 39, but it is sufficient enough for driving small headphones and speakers [7]. Also, an increase in gain only leads to an increase in noise and distortion in the output signal. The audio amplifier circuit designed using LM386-integrated circuit is a low-power circuit that can supply a maximum power of 1 W. This circuit is also bulkier because the LM386 works with large electrolytic capacitors, which causes it to be prone to distortions with duration of use. We have used TDA2822 IC to make the overall circuit smaller and more compact. Also, unlike the TDA2822 dual-channel IC, LM386 is a mono-audio amplifier [8]. Therefore, we can use only one of channel since the audio amplifier only able to amplify one signal at a time, and in our circuit, we can use two signals as input for amplification simultaneously.

3 Circuit Design

(See Table 1)

The amplifier circuit is powered by a battery of 6 V. For filtering any unwanted ripples from the power source, the power supply is passed through a filter capacitor (C5) of 100 µF. A smart phone will be providing the audio input. It will be plugged with an audio jack of 3.5 mm to receive the audio signal. The 3.5 mm audio jack consists of three wires—one for the ground connection and the other two wires are used for the left and right channel. The two speakers are of 4 ohms impedance

Table 1 Components used to build the circuit

Name of the component	Specification
Semiconductors	IC1 TDA2822M dual-power amplifier
Resistors	R1, R2, R3, R4—100 Ω
Capacitors	C1 and C2—1000 µF, C3 and C4—0.1 µF, C5—100 µF, and C6, C7, C8, C9—10 µF
Connectors	J2, J3—2-pin screw terminal and J1—3-pin screw terminal
Speaker	0.65 W, 4 Ω/0.38 W, 8 Ω
IC socket	8-pin DIP package
Potentiometers	U2, U3—2 volume control potentiometers
Power supply	6 V DC source
Audio jack	3.5 mm audio jack

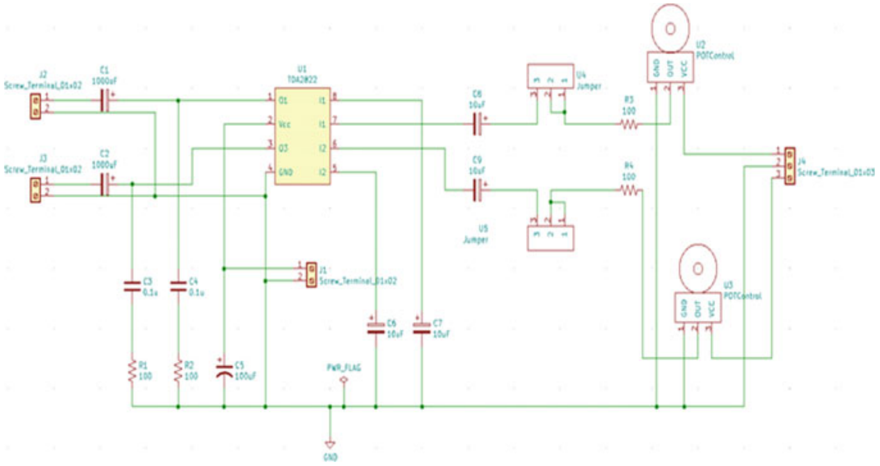


Fig. 1 TDA2822 stereo audio amplifier circuit

and are used at the output of the amplifier as load. The speakers are connected to the output pins of the TDA2822 at pins 1 and 3, and ground wires of the speakers are connected to the ground. TDA2822 being a dual-low-power audio amplifier has a structure of 8-pin plastic dual in-line package. Some of its features include low crossover distortion and low quiescent current with supply voltage range of 3–15 V. The TDA2822 is a dual-audio amplifier IC that means it has two op-amps inside its package, and because of their wide bandwidth gain, they are commonly used for audio amplification. 250 mW output power can be delivered by the two outputs. The TDA2822 IC finds its use in mini-radio, headphone amplifier, preamplifiers, portable audio systems, hearing aids, earphones, etc. It can be used in two modes: stereo mode and bridge mode. In this project, we will be using it in stereo mode (Fig. 1).

The left and right speakers are connected to the output pins 1 and 3 of the TDA2822 IC through electrolytic capacitors C1 and C2, respectively. The inverting input pins 5 and pin 8 are connected to the ground via the filter capacitors C6 and C7. The non-inverting input pins (Pin 6 and pin 7) are connected to the potentiometers through electrolytic capacitors C8 and C9. Any DC component from the amplifier to the load (speaker) can damage it. Further, it also produces noise or distortion in the audio output. Therefore, to make sure that no DC component from amplifier IC passes to the output load, capacitors C8 and C9 are connected. The potentiometers U2 and U3 act as left and right volume controls for both the speakers. Pin 4 is connected to the ground, and Pin 2 is connected to the DC supply (V_{cc}). Across V_{cc} and ground pins, a filter capacitor C2 is connected. The ratio of the output divided by the input is called as amplifier gain. It indicates how much change has occurred in the output signal for a certain change in the input signal. The gain of the TDA2822M is high, specifically 39 dB give or take. This means to drive it to its maximum power you only need a small signal, and no preamplifier would be needed. This makes it ideal for plugging into portable devices such as the headphone output of music players and mobile phones. The output power of the IC is dependent on the input supply voltage and output load. We saw that at 6 V power supply, a 4 Ω speaker gave an output of 0.65 watts, and an 8 Ω speaker gave an output of 0.38 watts. The jumpers are necessary to carry the signal at both high and low frequencies. Basically, they allow the user to switch between single-channel and dual-channel operation of the circuit (Figs. 2 and 3).

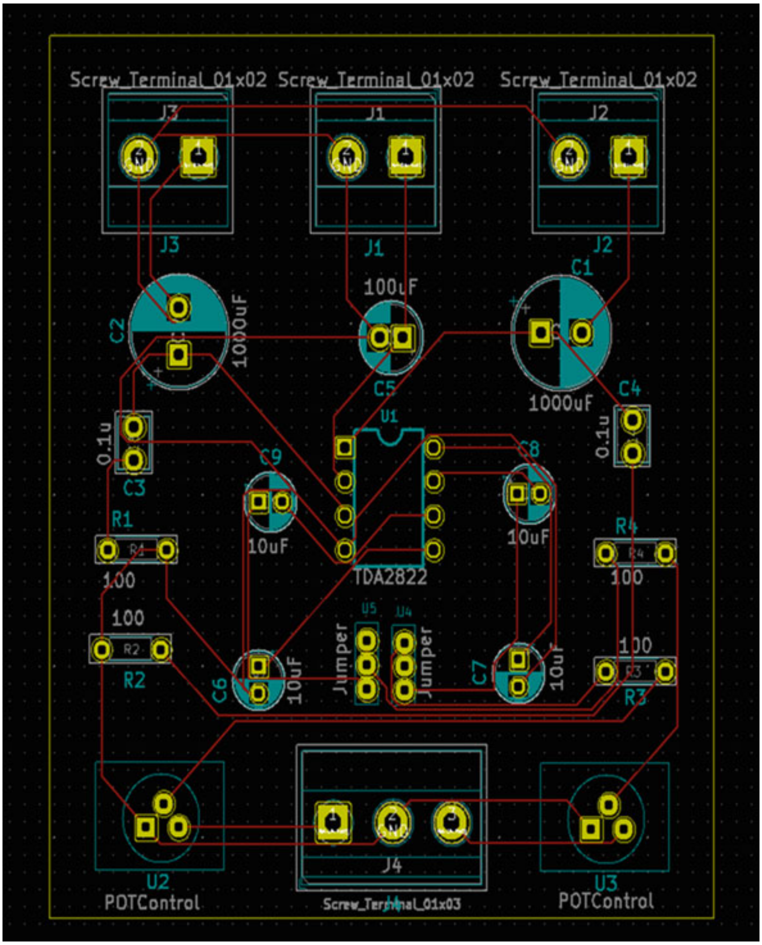


Fig. 2 PCB layout of amplifier circuit in Ki-CAD

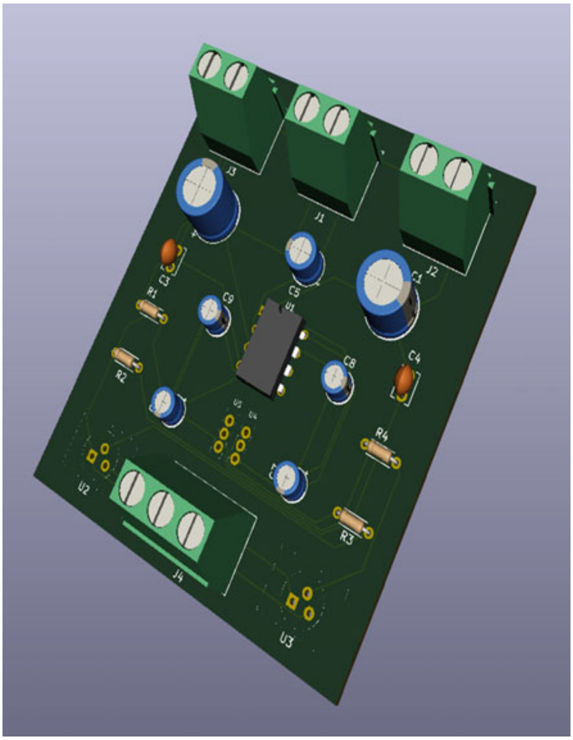


Fig. 3 PCB model of stereo audio amplifier board using TDA2822

4 Conclusion

The objective of the project was to create an audio amplifier circuit with less components and minimum noise distortion in the output signal. This circuit performs well with music having more of vocals and acoustic instrument sequences. Therefore, this IC is more suitable for driving smaller loudspeakers. It can be used as a driver for headphones when no particular performance is needed. The circuit was successfully designed using minimum peripheral components, and a fairly good audio signal output can be obtained without much noise or distortion.

References

1. Chen WT, Jones RC (2001) Concepts and design of filterless class-D audio amplifiers. Texas Instrum Tech J 18(2)
2. Pillonnet G, Abouchi N, Cellier R, Nagari A (2009) A 0.01%THD, 70dB PSRR single ended class D using variable hysteresis control for headphone amplifiers. ISCAS, Taipei, Taiwan, pp 1181–1184
3. Salahddine K, Qjidaa H (2009) Audio power amplifier solutions for new wireless phones. IJCSNS Int J Comput Sci Netw Secur 9(1)

4. Agung NR, Takagi S, Fujii N (2002) Improving the immunity to substrate noise using active noise cancellation circuit in mixed-signal integrated circuits. In: Asia-Pacific conference on circuits and systems, vol 1, pp 135–140. https://doi.org/10.1109/APCCAS.2002.1114923
5. Arefin A, Rahman U, Razzak M (2012) Design & implementation of a 25-watt audio power amplifier. https://doi.org/10.13140/2.1.3224.0961
6. Van der Zee RA, van Tuijl EA (1999) A power-efficient audio amplifier combining switching and linear techniques. IEEE J Solid-State Circuits 34:985–987
7. Salahddine K, Jalal L, Hajji (2012) Design a new architecure of audio amplifiers class-D for a hardware mobile systems. IJCSA Int J Comput Sci Appl 1(1):6–11
8. Chiesi L, Lorenzani E, Franceschini G, Bellini A (2006) Modeling of a parallel hybrid power audio amplifier. In: Proceeding IEEE industrial electronics conference, pp 1775–1780

Chapter 16
CAN Protocol-Based Modeling of Vehicle Comfort System

Anagha Liladhar Chaudhari and Suchitra A. Khoje

1 Introduction

The automatic climate control system present in this paper is developed using MATLAB-based model of vehicle and CAN protocol. A commonly used protocol for serial communication is a controlled area network (CAN) bus protocol. The devices within a vehicle can communicate with each other using CAN without host computer [1].

The most advanced air conditioning systems in cars are automatic climate control system. ACCS can effectively control the cabin temperature and humidity levels. In climate control, the cabin temperature of your choice can be set by you. The system controls it without the outside air temperature, and humidity is taken into account. Some advanced systems use automatic re-circulation mode, and so, they offer dual-zone climate control.

The ACCS provides good climate for the occupants. The air distribution, air flow, and temperature inside the car cabin are automatically gets controlled. The air circulation and fan speed also get controlled by climate control system [2].

Automatic climate control is a system that allows a driver to set a preferred temperature and then not worry about re-adjusting the heat or air conditioning, thanks to in-vehicle sensors.

The automatic climate control enables the vehicle occupant to set temperature within the vehicle to the specified value. This is its advantage. Then, the temperature is regulated by the automatic climate control system; the system then decides how to maintain the current environment.

A. L. Chaudhari (✉) · S. A. Khoje
Department of Electronics and Communication, MIT World Peace University, Kothrud, Pune, India

S. A. Khoje
e-mail: suchitra.khoje@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_16

Regardless of the outside weather conditions, the maintenance of the temperature of passenger compartment and humidity is offered by automatic temperature control, at a preset level. During cold weather, the passenger compartment is continuously gets heated to the particular temperature level, and then, system automatically maintains it.

A manual air conditioning system stays on at the cooling and blower setting you can select and keeps on blowing at that pace until you switch it lower. ACCS automatically keeps your vehicle at a specific temperature you select.

Automatic climate control measures things like sun heating, temperature of outside air, humidity inside vehicle, and direction and speed of your vehicle. This information is used to work ACCS automatically and help to keep your vehicle's interior at your preferred temperature [3].

2 Literature Survey

2.1 Literature Survey

A CAN protocol-based vehicle control system is designed to improve the driver and vehicle interface. CAN focused on the implementation of digital driving system and development of a semi-autonomous vehicle. The ARM microcontroller-based data collection system which uses analog-to-digital converter to convert all the control data, collected from the devices connected in the system, in analog-to-digital format and display that through LCD. The module used for communication in this paper has efficient data transfer because of an embedded networking by CAN. Engine temperature, power seat, power window, vehicle speed, antilock braking system, etc., provide the feedback about vehicle conditions. Digital-driving behavior means the joint mechanism development for vehicle to establish a control and decision-making framework between driver and vehicle [4].

In "Vehicle control system implementation using CAN protocol," it is mentioned that controlled area network can enhance the speed, security, performance, and utility of a system as it uses CAN as device. Two nodes of CAN bus are connected to each other by 2 Mbps. Microcontroller A connects voltage and temperature at NODE A through CAN. Microcontroller B is at NODE B and connected through CAN controller. Microcontroller B is connected to an IR sensor and machine control to exchange the automatic info about automobile. NODE A sensors sense the captured information change, i.e., temperature change and passes it over the CAN connected at NODE A. This data of CAN are transferred to NODE B of CAN, and then, LCD connected to microcontroller of NODE B displays it. The increase or decrease the speed of motor based on intensity of light is sensed by IR sensor [5].

In "The FTT-CAN protocol: why and how," with respect to time-triggered communication, how the CAN protocol differs from other protocols is described in this paper.

Communication overhead is getting reduced with high efficiency, and flexibility in the time-triggered traffic is supported by CAN reduces [6].

2.2 Controlled Area Network (CAN)

Controlled area network (CAN) is a commonly used serial communication bus protocol. International Standards Organization (ISO) defines it. CAN protocol connects devices inside the car as shown in Fig. 1 [1]. Serial communication is used when data is transferred from one network or point to another point is done in duplex mode. CAN protocol replaces the complex wiring by a two-wire bus. CAN protocol was developed for automotive industry. The electrical interference was reduced by the use of CAN. The transferred signal in the network was interfered by noise. An error checking and correcting mechanism was introduced by CAN protocol, which was very efficient in transferring the correct and authenticate data over the network.

In industries, including manufacturing, medical, and building automation, CAN is very popular in communication. Information is passed between devices on a network is explained by the CAN communications protocol explains the information passing between the on-network devices, ISO-11898: 2003. OSI model was followed by CAN that is defined in terms of different layers as shown in Fig. 2 [1]. OSI model consist of seven layers. Out of which, physical layer and data link layer are used in CAN communication.

Data link layer and physical layer of OSI model are as shown Fig. 3. The actual communication between connected devices was explained by the physical layer of the OSI model CAN. The connection between two nodes is done through a physical wire. The bottom two layers out of the seven layers of OSI/ISO model as the data

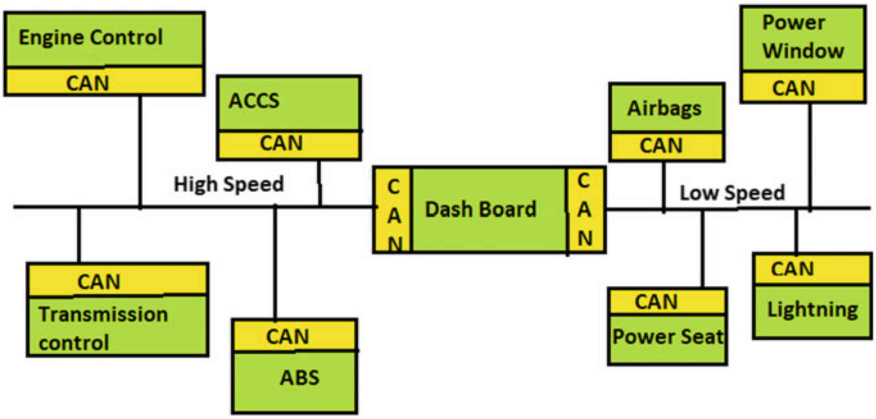


Fig. 1 Devices connected to each other with CAN protocol

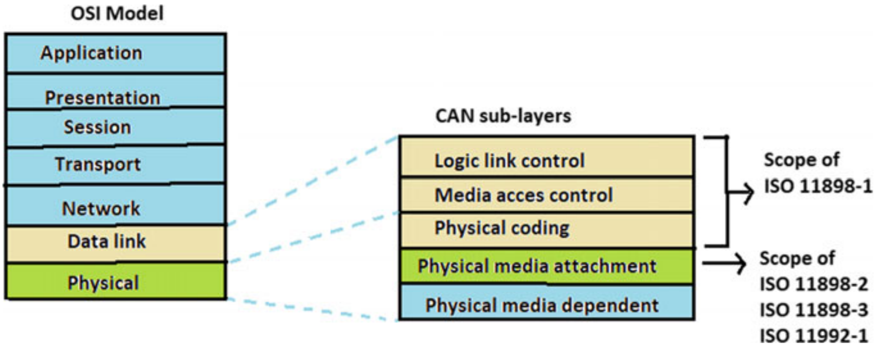


Fig. 2 OSI model

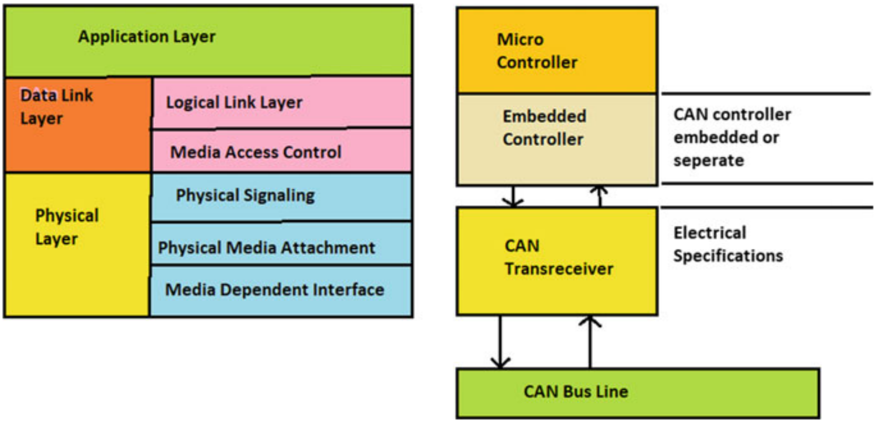


Fig. 3 OSI layers used in CAN

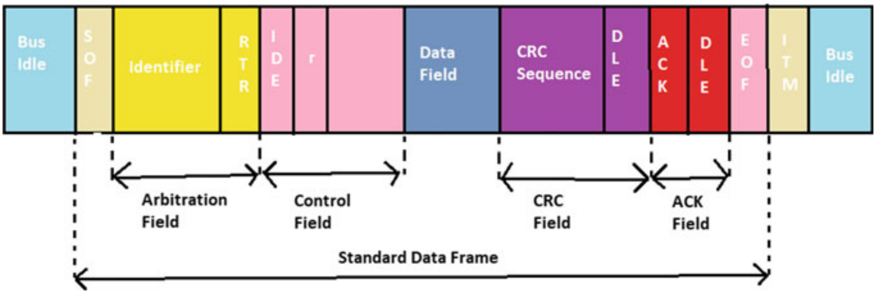


Fig. 4 CAN standard frame format

link layer and physical layer were defined by the ISO 11898 architecture. Figure 4 shows in detail the usage of data link layer and physical layer in CAN [1].

2.3 CAN Standard Frame Format

There are message frames present in CAN are of four different types:

- Data Frame—data frame is used for data transfer from one transmitting node to one or numerous receiving nodes.
- Remote Frame—remote frame is used for making requesting source node data. Data frame containing the requested data was consequently followed by remote frame.
- Error Frame—error frame is used for detection of an error condition. Error can occur at any time during a remote frame or data frame transmission, sender or receiver, any bus participant.
- Overload Frame—when a node can request a delay between the transmissions of two remote or data frames, then overload frame is introduced in the transmission. This also means that the overload frame occurrence is possible only between remote or data frame transmissions. There is minimum of three-bit times distance between consecutive frames [7].

As per definition, a CAN protocol remote or data frame has the following components:

- SOF (Start of Frame)—start of frame will indicate the starting of remote and data frames.
- Arbitration Field—arbitration field will provide the information about remote transmission request (RTR) bit and message ID, which is used for distinguishing between data and remote and data frames present in the message.
- Control Field—control field is used for determination of message ID length and data size in the CAN message.
- Data Field—data field is used for introducing the actual data in message (This field is applied only to a data frame and not a remote frame)
- CRC Field—cyclic redundancy check is the checksum. It is introduced in CAN frame. Checksum from the transmitted bits is calculated by the transmitter, and the result is provided within the frame in the CRC field.
- EOF (End of Frame)—end of frame will indicate the end of data and remote frame. [8]

3 Methodology

General flow of proposed system is as shown in Fig. 5. In our proposed system, CAN protocol is used for communication. So, all the models are created using

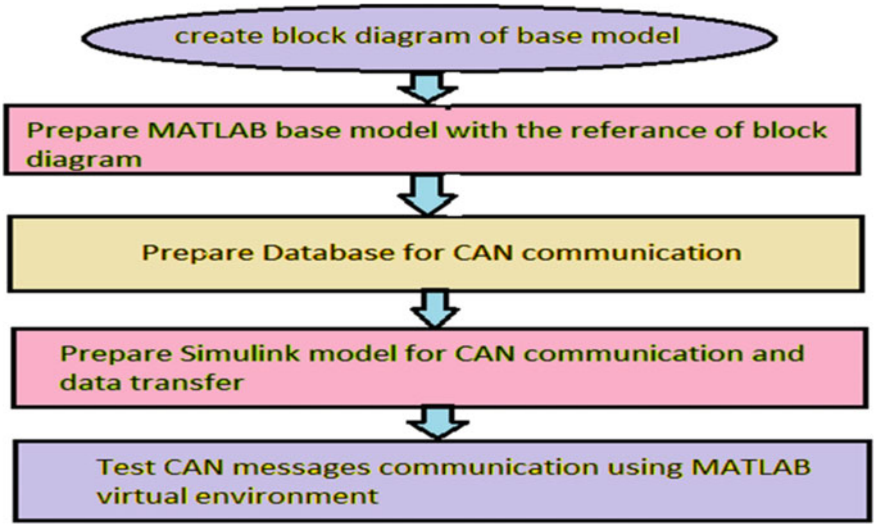

create block diagram of base model
Prepare MATLAB base model with the referance of block diagram
Prepare Database for CAN communication
Prepare Simulink model for CAN communication and data transfer
Test CAN messages communication using MATLAB virtual environment

Fig. 5 General flow of proposed system

MATLAB SIMULINK. In CAN, communication database creation is important, so in this proposed system, detailed explanation about database creation is also given. Block diagram of system is prepared; as per the blocks, internal mechanism of each block is finalized. The HVAC algorithm of the proposed system is implemented in the MATLAB base model. To set up, the CAN communication using created database target model is designed, and testing of model is done, with the help of MATLAB virtual environment. Detailed functionality of all the steps mentioned in general flow is as follows:

3.1 Block Diagram of System

The model of the proposed system is based on the block diagram shown in Fig. 6 is implemented using MATLAB. There are four major blocks in the system; out of which, first block allows user to select HVAC mode; its output selects the internal HVAC mode. According to that mode blower speed and blower opening blocks produce their respective outputs. When driver selects HVAC mode, this will activate the internal CAN bus by taking feedback from temperature sensor, fitted inside the vehicle, and sets internal mode as per temperature reading. There are total seven internal HVAC modes are provided; out of these, three are for controlling AC, three for heater, and one is comfortable mode. Comfortable mode is turned on by either user or any other mismatch in setting happens. It is a default mode of HVAC system. In blower speed block, how much amount of air should be blown inside the vehicle is decided and whether that air is hot or cold. For heater, hot air should blow and

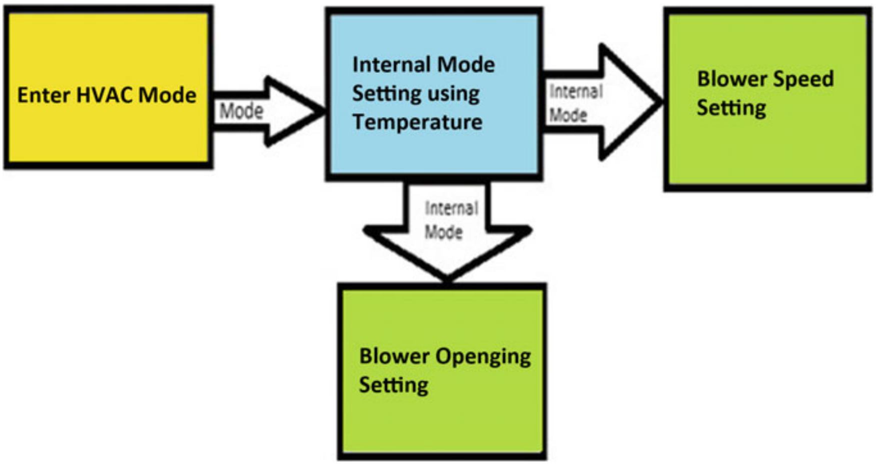


Fig. 6 Block diagram of MATLAB base model

cold for AC. Next block is blower opening; in this block, flap angle of the blower is decided.

3.2 MATLAB Base Model

As explained in the block diagram, vehicle occupant has facility to select between heater and AC. After getting the input from user, the internal temperature value is fetched from the sensors for driving further blocks, i.e., block number 2 which will select the internal HVAC mode. The temperature fetching from sensor will be done when we download the designed model in the test vehicle. In our proposed system, for the simulation and testing purpose, constants are used. Slider gain block used here is for checking all the conditions mentioned and for varying the constant values in the model. The use of slider gain block makes the model user-friendly. The MATLAB base model shown in Fig. 7 is designed in the MATLAB state-flow, consists of state-flow chart, constant block, slider gain block, and display blocks to display output at every point of the model [9].

The temperature ranges which will further decide the internal HVAC mode are explained in Table 1. Once the internal mode is decided, further blocks, i.e., block number 3 and 4 are driven. The block number 3 will decide the speed of blower and whether the air blown out of the blower should be hot or cold. Block number 4 will decide at what angle the flap of blower should remain open inside the vehicle. Inside the block number 2, there are two main states first will check whether the internal vehicle temperature is above 18 °C or not. If the ambient temperature is above 18 °C, then only, it will allow the heater or AC to be ON; otherwise, it will kept OFF. Second state block will decide the internal HVAC mode. Here, total 7

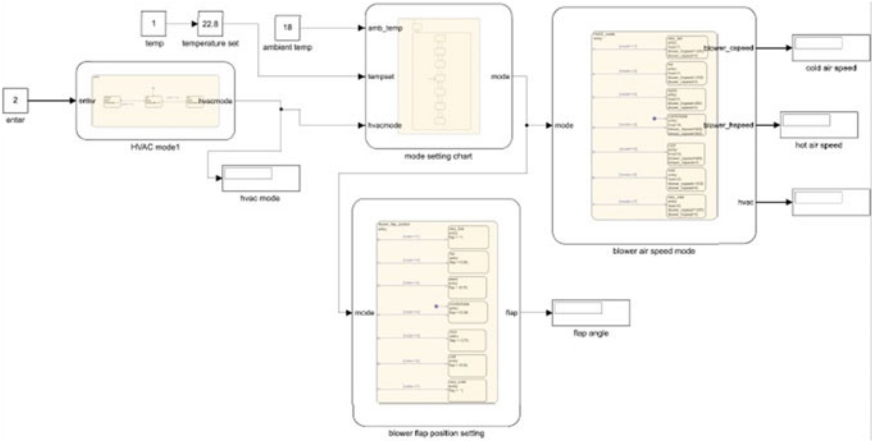


Fig. 7 MATLAB model

Table 1 Internal temperature ranges

Mode	Temperature range
Very hot	>28
Hot	26–28
Warm	24–26
Comfortable	22–24
Cool	20–22
Cold	18–20
Very cold	<18

modes are present internally, 3 are for heater, and 3 are for AC, and 1 is comfortable mode. In comfortable mode, both the AC and heater will be on [10].

Depending on this, internal HVAC mode blower speed and flap opening angle are being decided by the state-flow chart. There are two types of air blown inside the vehicle. Suppose heater is selected by user, then depending on temperature among three internal heater modes, one mode gets selected.

As there are two types of air to be blown inside the vehicle compartment, depending on the mode selected by block 2, blower speed is selected. Table 2 gives blower speed for each mode. The working of this block depends on the principle that as blower speed is high, more amount of air will be blown out. Here, air speed is taken in cubic feet per meter (CFM). When comfortable mode is selected, then both type of air is blown inside the vehicle.

Block 3 will give flap angle opening setting. Table 3 shows the flap angle for particular mode. This block works on the principle that when the flap angle is more, more amount of air will be blown out of the blower, and it will take control of atmosphere easily.

Table 2 Blower speed

Mode	Blower speed
Very hot	1270
Hot	1100
Warm	900
Comfortable	900, 900
Cool	900
Cold	1100
Very cold	1270

Table 3 Blower flap angle

Mode	Blower flap angle
Very hot	1
Hot	0.84
Warm	0.70
Comfortable	0.50
Cool	0.70
Cold	0.84
Very cold	1

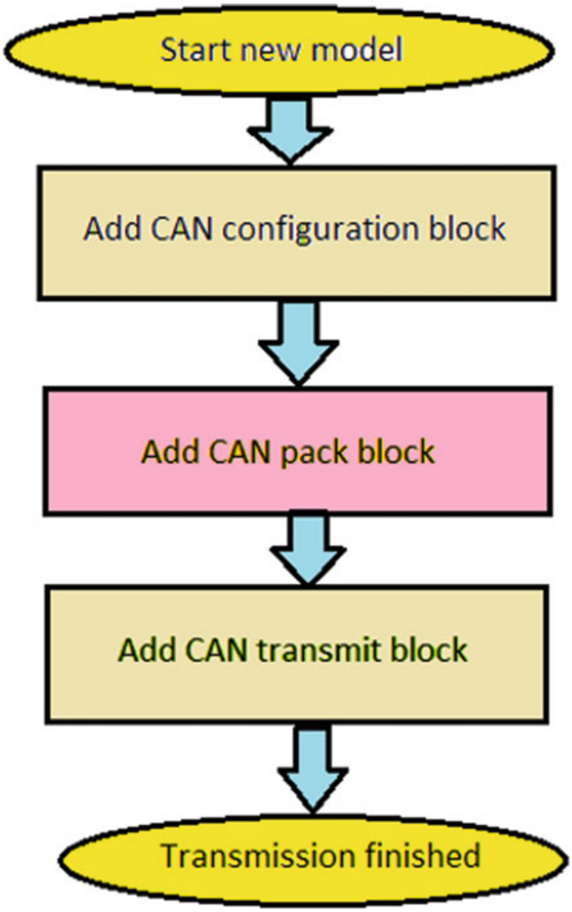
3.3 Target Model of System

Our proposed system is based on the CAN communication, so to set this communication, target model is created. Vehicle network toolbox in the MATLAB is used to create the target model of the system. Vehicle network toolbox is the single environment capable of handling all CAN traffic. Vehicle network toolbox directly connects MATLAB SIMULINK to CAN network because of which real-time simulation is possible, being capable of connecting virtual hardware with virtual networks. Vehicle Network Toolbox supports CANdb database file and allows loading and replying messages through it. Vehicle network toolbox also supports XCP protocol over CAN.

Transmission and reception of the messages over CAN bus are routed through this model. Model creation workflow is as follows. Figure 8 shows the messages transmission workflow, and Fig. 9 shows the messages reception workflow. Model used to transmit the messages is created by following transmission workflow, and using reception workflow model to receive messages over, CAN bus is created. Models are created according to the workflow shown in Figs. 10 and 11; this shows target model of the system.

The target model contains the CAN configuration, CAN transmit, and CAN receive blocks that are packed and unpacked using the CAN pack and CAN unpack blocks from vehicle network toolbox. The target model receives CAN messages through channel of virtual hardware from host model. The model transmits CAN

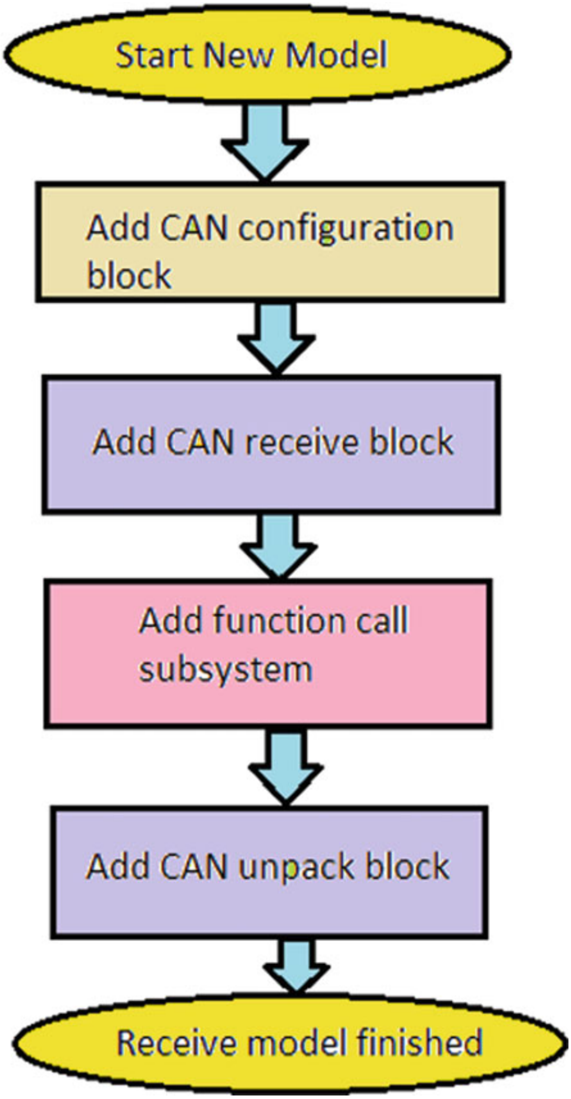
Fig. 8 Workflow of message transmission



messages using channel of virtual hardware over the CAN bus to host model and also to the other devises connected to the bus.

CAN unpack block is used for unpacking the signals inside the messages and allows for further transmission and packing of that signals to create new message. These created new messages were transmitted over the CAN channel. Figure 11 shows the functional subsystem of the model. All the packed signals are then transmitted over the channel to host model. For data transmission, there are two options either we can send signal periodically or event based. Also, select the devices for data transmission whether transfer it over channel 1 or channel 2. There should be synchronization between host and target. Both have to send data over same channel. To run this model successfully, the target model configuration settings must match to the host model configuration settings.

Fig. 9 Workflow of message reception



3.4 Creating CAN Database

After designing the base model, according to its functionality, we required in our proposed system that CAN database is created using CANDB++ editor. Figure 12 shows the general flowchart for the creation of database. Flowchart gives the exact idea about the steps involved in the process of creation of database. Detailed steps are as follows, there are few steps to create database from scratch.

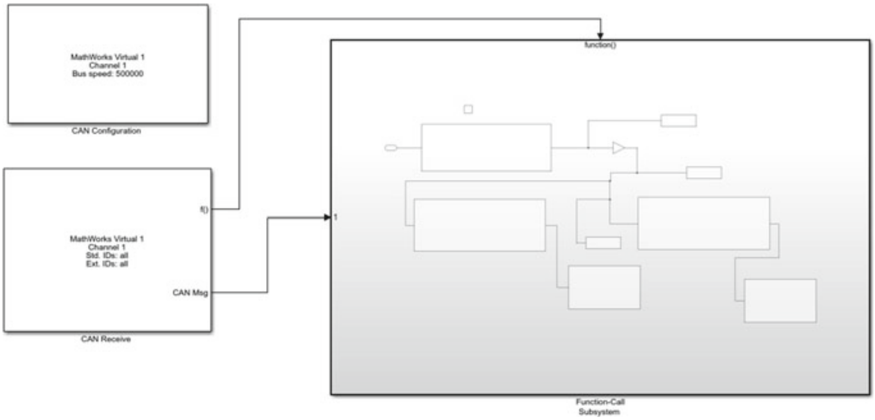


Fig. 10 Target model

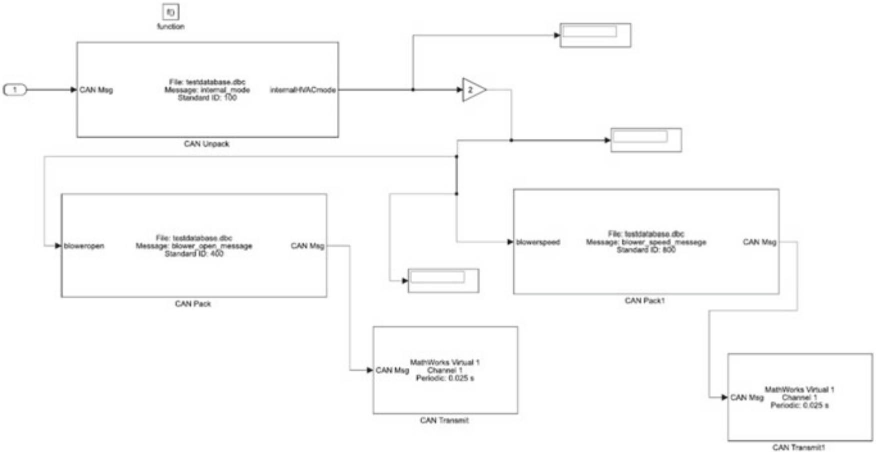


Fig. 11 Function subsystem of target model

(1) To create new database file, in file option click on menu bar, then click on create database.
(2) Select the template, appropriate for database.
(3) Select the location to store the database.
(4) Give proper file name and save the file.
(5) Now, create the signals, by right clicking on signals.
(6) Rename the signal name. Describe signal length in bits and its minimum and maximum values with unit.
(7) The secondary task bar next to menu bar has icon for creation of value table.
(8) By using this icon, create value table which contains all the values that signal can hold.

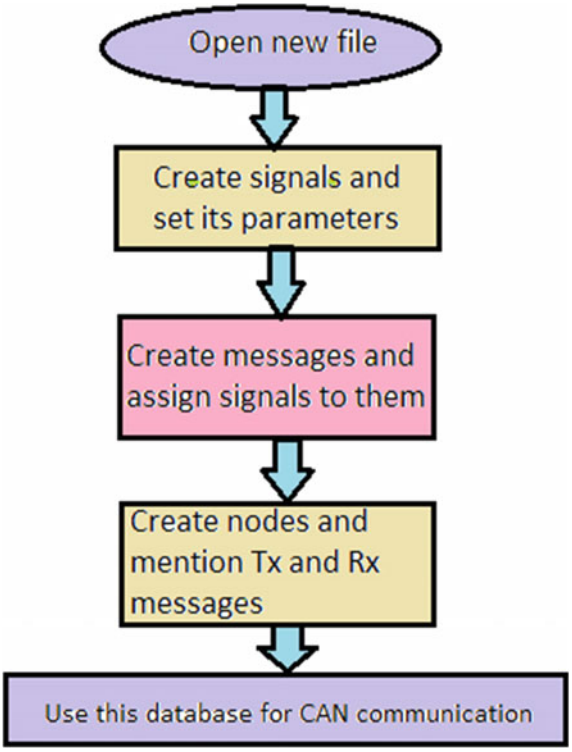
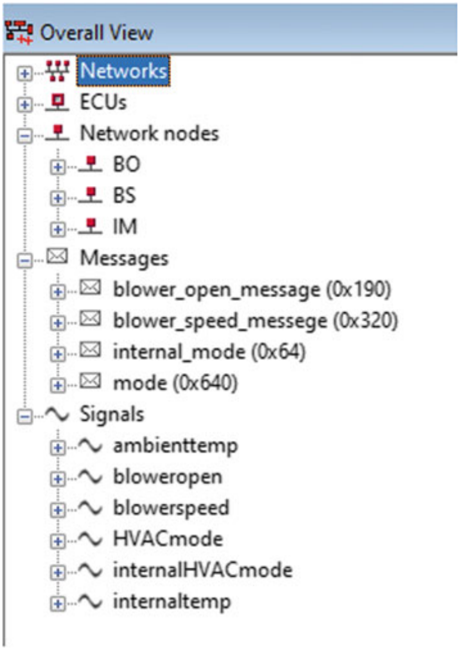

Open new file
Create signals and set its parameters
Create messages and assign signals to them
Create nodes and mention Tx and Rx messages
Use this database for CAN communication

Fig. 12 General database creation flowchart

(9) Then, assign that value table to create signal by editing it. Similarly, all required signals are created.
(10) Then, messages are created, by right clicking on it.
(11) Id of the message can also be defined.
(12) Rename the message name. Add signals to message, two or more signals can also be added. After creating messages, nodes are created.
(13) Network nodes are created by right clicking on it. Rename it as per definition.
(14) Define the signals mapped by that node, transmitted messages and received messages which are mapped by that node is to be defined.
(15) After defining that node automatically gets assigned to the selected messages.
(16) In the node definition, address of the node can also be selected.

In CAN communication, database is the base of all communication and data transfer over the CAN bus. By following above steps, database is created using CANDB++ editor. The created CAN database is having specifications shown below. The overall view of database contains information related to CAN communication, its components like nodes, messages, signals on which communication is based on. Figure 13 gives the overall view of the created database. This is the overall view of CAN database. This file of database is stored in .dbc file format, and for further communication, this file is used.

Fig. 13 Overall view of database



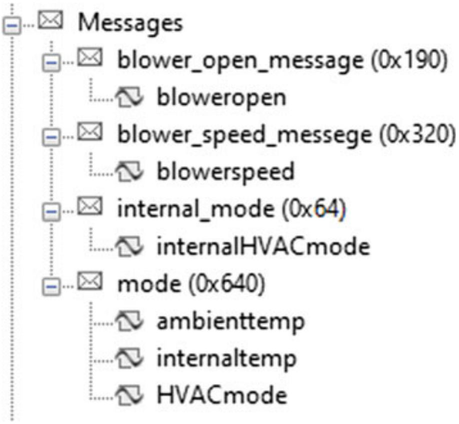
Overall View

- Networks
- ECUs
- Network nodes
  - BO
  - BS
  - IM
- Messages
  - blower_open_message (0x190)
  - blower_speed_messege (0x320)
  - internal_mode (0x64)
  - mode (0x640)
- Signals
  - ambienttemp
  - bloweropen
  - blowerspeed
  - HVACmode
  - internalHVACmode
  - internaltemp

Messages in database are shown in Fig. 14. There are four messages in the database, mode, internal mode, blower speed message, and blower open message. Figure shows them along with their signals.

Details of the nodes are shown in Fig. 15. There are three nodes, internal mode select, blower speed, and blower open. Figure gives information about mapped transmitted signals, mapped Rx signals, transmitted message, received message related to that node.

Fig. 14 Messages description in database



- Messages
  - blower_open_message (0x190)
    - bloweropen
  - blower_speed_messege (0x320)
    - blowerspeed
  - internal_mode (0x64)
    - internalHVACmode
  - mode (0x640)
    - ambienttemp
    - internaltemp
    - HVACmode

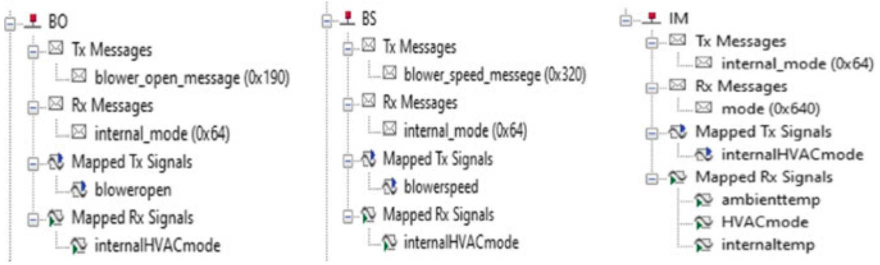


Fig. 15 Nodes description in database

3.5 Testing Using MATLAB

In our paper, one system inside the vehicle, i.e., ACCA, is taken into consideration, and its modeling is discussed in detail. So, the testing of this proposed model is done in MATLAB virtual environment. As the proposed system is completely based on the CAN messages transmission and reception, MATLAB 2021a version is used to carry out all the experimentation before downloading the model into actual vehicle. Testing of the proposed system is important because it will give error check of designed model. For demonstration purpose, database exploration is done in MATLAB which includes the use of command provided by MATLAB.

4 Results

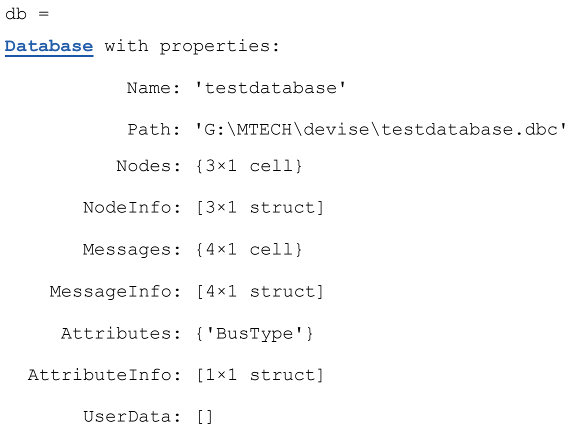
In our proposed system, vehicle network toolbox provided by MATLAB is used for sending, receiving, encoding, and decoding CAN messages. In order to validate the accuracy of this developed model and database, testing is done in virtual environment. To test the model, MATLAB provides the commands to explore the functionality of the database. To test the CAN communication, start CAN explorer in bag round, and once model simulation starts, messages get displayed on CAN explorer. The graph message distribution over time is plotted below. Also, the change in signal values with respect to time is observed, and its graph is plotted.

MATLAB commands with its associated output is mentioned below.

i. **Load Database**

```
db = canDatabase('testdatabase.dbc')
```

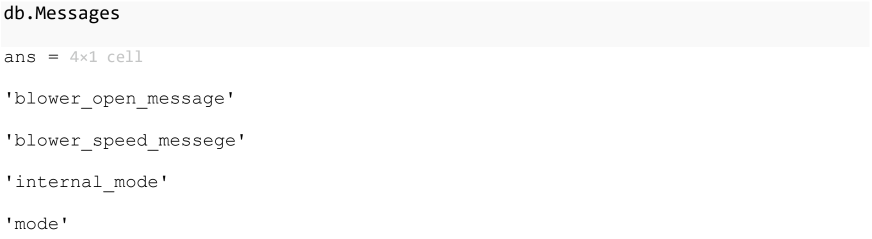
Load the database. Database is having following properties.



```
db =
Database with properties:
                Name: 'testdatabase'
                Path: 'G:\MTECH\devise\testdatabase.dbc'
               Nodes: {3×1 cell}
            NodeInfo: [3×1 struct]
            Messages: {4×1 cell}
         MessageInfo: [4×1 struct]
          Attributes: {'BusType'}
       AttributeInfo: [1×1 struct]
            UserData: []
```

It shows there are three nodes, four messages present in the database.

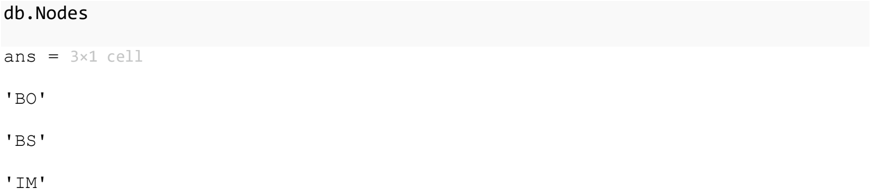
ii. Check messages present in Database



```
db.Messages
ans = 4×1 cell
'blower_open_message'
'blower_speed_messege'
'internal_mode'
'mode'
```

These are those for messages on which the functionality of model is dependent.

iii. Check nodes present in Database



```
db.Nodes
ans = 3×1 cell
'BO'
'BS'
'IM'
```

These are the nodes present to establish the communication between host and target. Nodes will take care of transmission and reception if messages properly. Nodes gives the detail information about mapped transmission signals, mapped reception signals, and signals to be transmitted by that node.

iv. Signals associated with messages

```
msgEngineInfo.Signals
```

```
struct with fields:

bloweropen: 0

blowerspeed: 900

internalHVACmode: 4

HVACmode: 0

internaltemp: 18

ambienttemp: 18
```

v. Checking signals in message

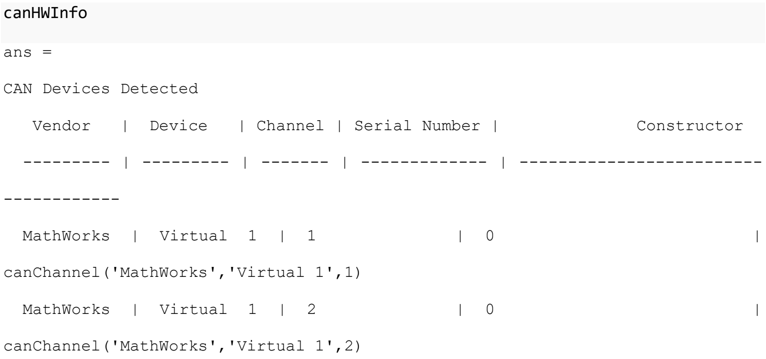
```
msgEngineInfo = canMessage(db, 'mode')

msgEngineInfo.Signals
```

```
msgEngineInfo =

  Message with properties:

   Message Identification

    ProtocolMode: 'CAN'

              ID: 1600

        Extended: 0

            Name: 'mode'

   Data Details

       Timestamp: 0

            Data: [0 0 0]

         Signals: [1×1 struct]

          Length: 3

   Protocol Flags

           Error: 0

          Remote: 0

   Other Information

        Database: [1×1 can.Database]

        UserData: []

ans = struct with fields:

        HVACmode: 0

    internaltemp: 18

     ambienttemp: 18
```

This gives all the information about message. In the form if parameters like Message ID, length, data and signals it holds. Similarly, we can get information about all the signals. Here, data field is empty, as it is testing in virtual environment.

To start the message transmission for establishment of communication between host engine control unit and model. The step is to check the hardware of our system.

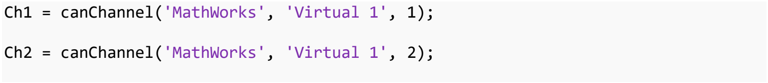
vi. Checking hardware connected to system



```
canHWInfo
ans =
CAN Devices Detected
   Vendor    |  Device    | Channel | Serial Number |               Constructor
  --------- | --------- | ------- | ------------- | -------------------------
------------
  MathWorks  |  Virtual  1  |  1               |  0                             |
canChannel('MathWorks','Virtual 1',1)
  MathWorks  |  Virtual  1  |  2               |  0                             |
canChannel('MathWorks','Virtual 1',2)
```

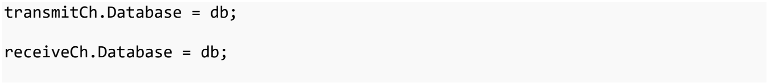
Assign this virtual channel for transmission and reception purpose, respectively.

vii. Create CAN channel for transmission and reception



```
Ch1 = canChannel('MathWorks', 'Virtual 1', 1);
Ch2 = canChannel('MathWorks', 'Virtual 1', 2);
```

These commands are used to allocate the channels for transmission and reception of virtual messages. Virtual channel 1 is used for transmission, and at the same time, virtual channel 2 is used for reception of messages.



```
transmitCh.Database = db;
receiveCh.Database = db;
```

Allocate database to the channel.

viii. Configure messages for transmission

```
transmitPeriodic(transmitCh, msg(1), 'On', 0.100);

transmitPeriodic(transmitCh, msg(2), 'On', 0.200);

transmitPeriodic(transmitCh, msg(3), 'On', 0.500);

transmitPeriodic(transmitCh, msg(4), 'On', 0.400);
```

Decide the periodic frequency of each of the message in the database.

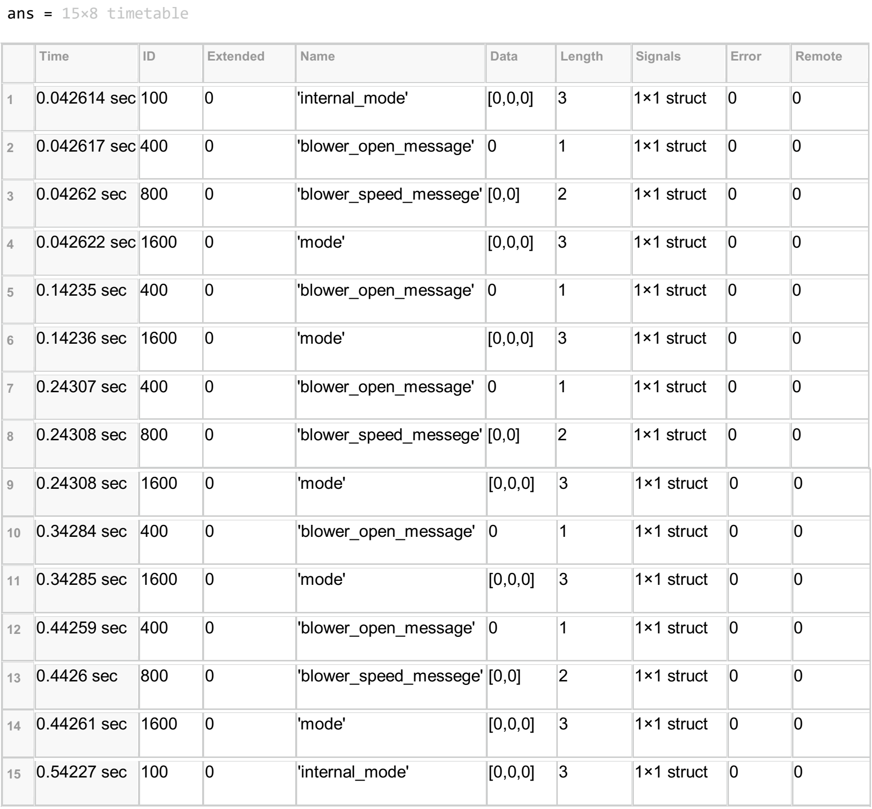
```
start(receiveCh);

start(transmitCh);

pause(2);
```

Start transmission and reception channel simultaneously. And wait for 2 s so that it can load some messages.

ix. Receive messages

```
stop(transmitCh);

stop(receiveCh);

msgRx = receive(receiveCh, Inf, 'OutputFormat', 'timetable');

msgRx(1:15, :)
```

Then, stop both the channels for receiving messages and display the messages transmitted and received over the CAN channel by the time it is on.



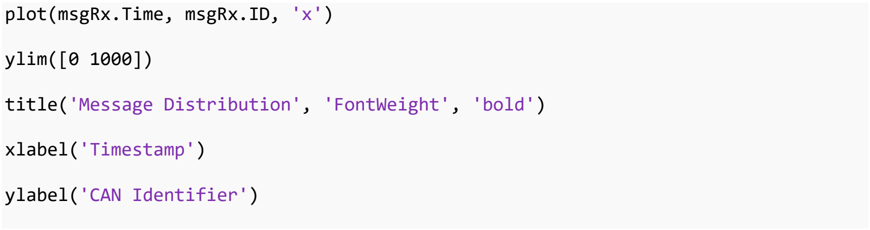
ans = 15×8 timetable

	Time	ID	Extended	Name	Data	Length	Signals	Error	Remote
1	0.042614 sec	100	0	'internal_mode'	[0,0,0]	3	1×1 struct	0	0
2	0.042617 sec	400	0	'blower_open_message'	0	1	1×1 struct	0	0
3	0.04262 sec	800	0	'blower_speed_messege'	[0,0]	2	1×1 struct	0	0
4	0.042622 sec	1600	0	'mode'	[0,0,0]	3	1×1 struct	0	0
5	0.14235 sec	400	0	'blower_open_message'	0	1	1×1 struct	0	0
6	0.14236 sec	1600	0	'mode'	[0,0,0]	3	1×1 struct	0	0
7	0.24307 sec	400	0	'blower_open_message'	0	1	1×1 struct	0	0
8	0.24308 sec	800	0	'blower_speed_messege'	[0,0]	2	1×1 struct	0	0
9	0.24308 sec	1600	0	'mode'	[0,0,0]	3	1×1 struct	0	0
10	0.34284 sec	400	0	'blower_open_message'	0	1	1×1 struct	0	0
11	0.34285 sec	1600	0	'mode'	[0,0,0]	3	1×1 struct	0	0
12	0.44259 sec	400	0	'blower_open_message'	0	1	1×1 struct	0	0
13	0.4426 sec	800	0	'blower_speed_messege'	[0,0]	2	1×1 struct	0	0
14	0.44261 sec	1600	0	'mode'	[0,0,0]	3	1×1 struct	0	0
15	0.54227 sec	100	0	'internal_mode'	[0,0,0]	3	1×1 struct	0	0

Table gives all the information about message received. It is ID, name, data, length, error along with the timestamp.

Now, to get knowledge about message distribution over time, graph is plotted.

x. Analyze the transmitted messages



```
plot(msgRx.Time, msgRx.ID, 'x')

ylim([0 1000])

title('Message Distribution', 'FontWeight', 'bold')

xlabel('Timestamp')

ylabel('CAN Identifier')
```

For experimentation purpose, some constant values of signal are assumed to check the smooth working of data field in messages (Fig. 16).

xi. Modify signals value in the message and plot updated signal data

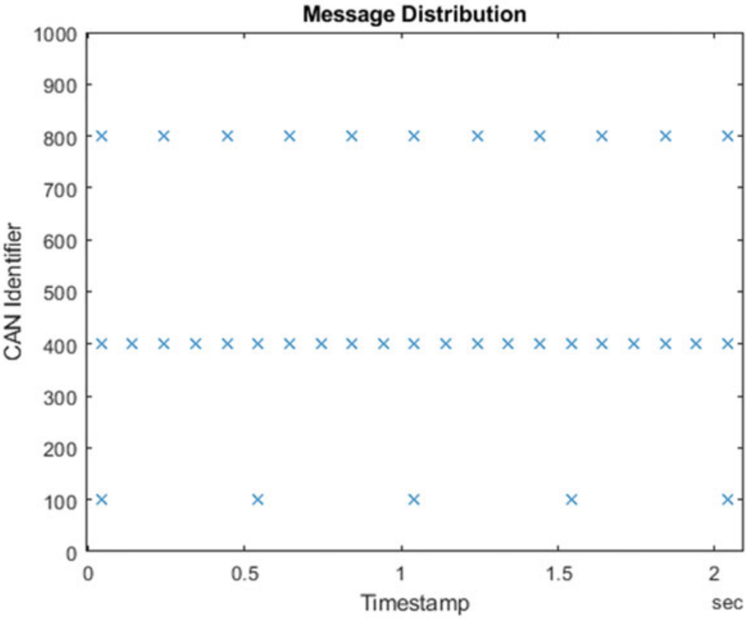
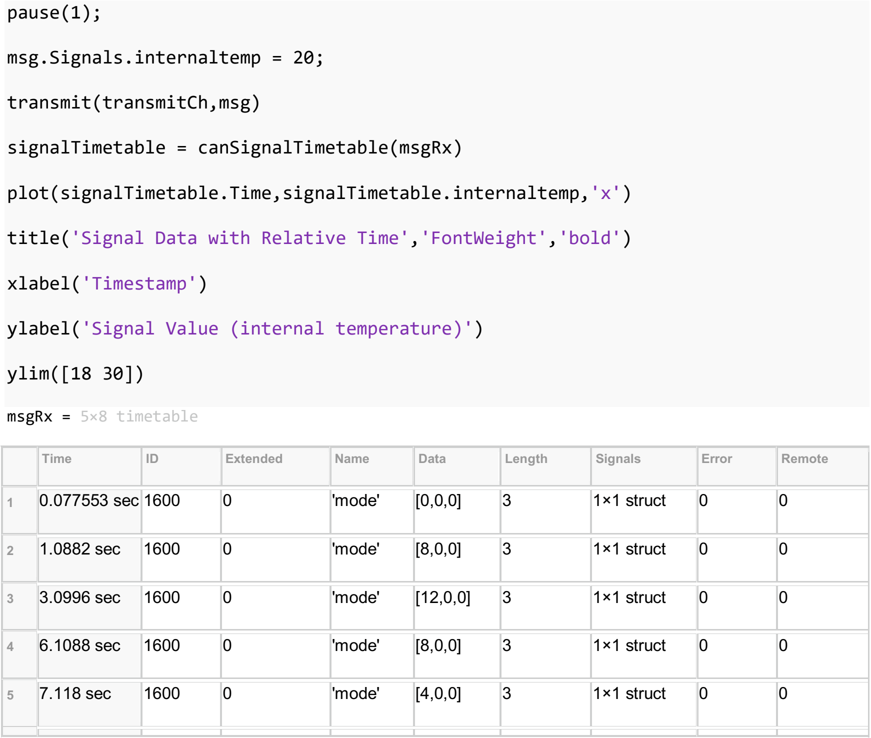


Fig. 16 Message distribution

```
msg.Signals.internaltemp = 18;

transmit(transmitCh,msg)

pause(1);

msg.Signals.internaltemp = 22;

transmit(transmitCh,msg)

pause(2);

msg.Signals.internaltemp = 24;

transmit(transmitCh,msg)

pause(3);

msg.Signals.internaltemp = 22;

transmit(transmitCh,msg)
```

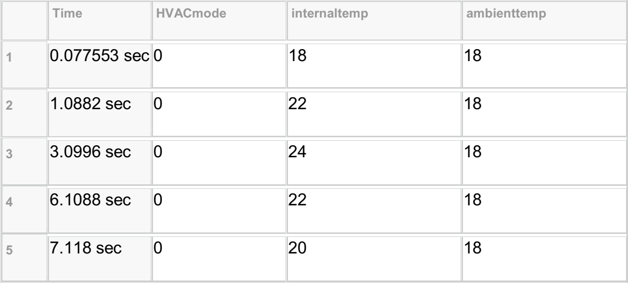


```
pause(1);

msg.Signals.internaltemp = 20;

transmit(transmitCh,msg)

signalTimetable = canSignalTimetable(msgRx)

plot(signalTimetable.Time,signalTimetable.internaltemp,'x')

title('Signal Data with Relative Time','FontWeight','bold')

xlabel('Timestamp')

ylabel('Signal Value (internal temperature)')

ylim([18 30])
```

```
msgRx = 5×8 timetable
```

	Time	ID	Extended	Name	Data	Length	Signals	Error	Remote
1	0.077553 sec	1600	0	'mode'	[0,0,0]	3	1×1 struct	0	0
2	1.0882 sec	1600	0	'mode'	[8,0,0]	3	1×1 struct	0	0
3	3.0996 sec	1600	0	'mode'	[12,0,0]	3	1×1 struct	0	0
4	6.1088 sec	1600	0	'mode'	[8,0,0]	3	1×1 struct	0	0
5	7.118 sec	1600	0	'mode'	[4,0,0]	3	1×1 struct	0	0

After updating the values of signal, data field of the messages also gets updated. The above table shows only those messages which are updated.



	Time	HVACmode	internaltemp	ambienttemp
1	0.077553 sec	0	18	18
2	1.0882 sec	0	22	18
3	3.0996 sec	0	24	18
4	6.1088 sec	0	22	18
5	7.118 sec	0	20	18

Signal table shows updated values with timestamp. Graph of internal temperature values against timestamp is plotted. Graph shows signal value with relative time (Fig. 17).

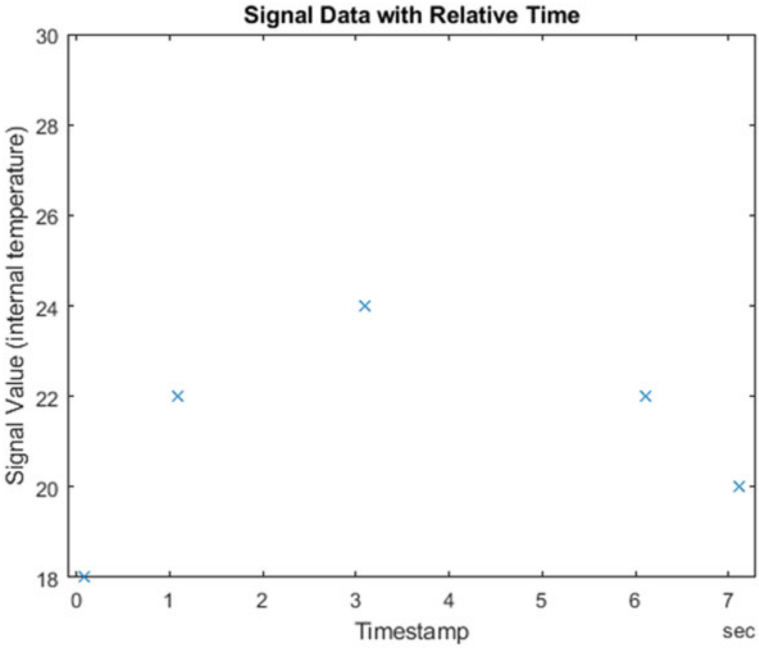


Fig. 17 Signal value with relative time

5 Conclusion

The customer's requirements about safety, convenience, comfort of the vehicles are increased daily. In addition to this, government policies and legislation for controlling pollution and reducing usage of fuel become strong day by day. So, car industry in association with ISO developed many modifications cum versions in electronic systems which lead to complex models to meet the entire customer's requirements. Complexity of this systems increases as it requires many dedicated single-line hard-wired connections for data exchange, and overall wiring increases the cost and physical size of system. In addition to this, the problem of fault diagnostics, repair, and reliability arises during processes, manufacturing, and service.

So, to address this issue, specifically, automatic climate control modeling is proposed in the paper. The model is created in MATLAB state-flow, and SIMULINK is used for creation of base and target model. Hard wiring issue between the microcontrollers and devises inside the vehicle reduces by CAN protocol. One system inside the vehicle, i.e., ACCS, is taken into consideration in this paper, and its modeling is discussed in detail.

There are many disadvantages of hard wired connections which include error detection and correction. In the original vehicle, many devises are connected to the engine control unit along with ACCS, but for simplicity, only one device is considered. The ACCS model for CAN transmission is developed using MATLAB vehicle network toolbox as it is the single environment capable of handling all CAN traffic. Vehicle network toolbox directly connects MATLAB SIMULINK to CAN network because of which real-time simulation is possible, being capable of connecting virtual hardware with virtual networks.

The problems faced by previous climate control systems were eliminated by automatic climate control system based on CAN protocol. It helps system to run in an efficient manner, reducing losses of data during transmission. The CAN provides self-diagnostics capability to prevent the system from malfunctioning. CAN protocol allows various systems connected using it to transmit signals simultaneously and individual system need not be waiting for the transmission till other system is going to complete its transmission.

Further, automotive systems modeling like engine control unit and antilock braking system, power ignition system modeling can also be possible using CAN protocol. Once all the systems are designed, the system becomes ready to download in the test vehicle and run on its own.

References

1. Chaithra Chandrashekar DBR, Vijay R (2014) Automatic car AC control using CAN protocol. 2: 2348–7968
2. Carbiketech.com on 5 April 5 2019. <https://carbiketech.com/automatic-climate-control-automatic-ac/>
3. my.chevrolet.com<https://my.chevrolet.com/how-to-support/getting-started/automatic-climate-control>
4. Singh VK, Archana K (2013) Implementation of CAN protocol in automobiles using advance embedded system. (IJETT) 4(10)
5. Vijayalakshmi S (2013) Vehicle control system implementation using CAN protocol. Int J Adv Res Electr, Electron Instrum Eng 2(6)
6. Almeida L (2002) The FTT-CAN protocol: why and how. IEEE Trans Ind Electron 49(6)
7. BOSCH CAN Handbook, 4th edn, (Volume-1 and Volume-2)
8. Ratcliff K An overview of control area network
9. mathworks.com/help/vnt
10. Wei KC, Dage GA (1995) An intelligent automotive climate control system. IEEE, pp 7803–2559

Chapter 17
Use of Blockchain for Securing Electronic Health Records During COVID-19 Pandemic

Nilima V. Pardakhe and V. M. Deshmukh

1 Introduction

Coronavirus-2019 (COVID-19) has begun to spread across the globe in the year 2020. Globally, the virus's massive spread rocked the health sector's underpinnings, which were ill-prepared and unable to react effectively. As a result, technological advances may aid in the restoration of human lives worldwide. Due to the high potential of pandemic risk, the World Health Organization (WHO) has advised that all nations develop a "Pandemic Plan." A Pandemic Plan is usually prepared under the WHO's pandemic phases and seeks to produce tangible outcomes in the early stages of pandemic management. The electronic health record (EHR) is a digital representation of a patient's real-time official health record that can be exchanged conveniently, securely, and instantaneously across various institutions and departments. It contains all the information necessary to get a patient's specifics, such as medical history, radiological pictures, diagnoses, medicines, vaccination dates, treatment plans, allergies, and test results. It is critical in health care since it facilitates contact with health archives to make patient treatment decisions. Due to the lack of standards and legislation around file sharing, EHR interoperability has become a critical problem for healthcare practitioners to address [1]. The confidentiality and security issues exacerbate the difficulties inherent in achieving interoperability during information exchange [2]. EHRs and other health information technology (HIT) organizations are disorganized due to insufficient communication morals across different EHRs, increased incorporation costs, short patient contribution in data distribution, and a lack of patient identification throughout health information exchanges (HIE).

N. V. Pardakhe (✉) · V. M. Deshmukh
CSE Department, P.R.M.I.T.&R, Badnera, Amravati, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_17

1.1 Electronic Health Records

The term “electronic health care” refers to the storage of patient data in a digital format. Patient-centered records enable data to be accessed remotely and at any time by any authorized user. Annual savings of more than $81 billion are attributed to the computerized health system (Hillestad et al., 2005). Electronic health care improves social and health outcomes while also minimizing medical mistakes.

The HIMSS focused on North America, Asia Pacific, the United Kingdom, and the Middle East. The HIMSS contains 72,000 individuals and 630 organizations. The main aim of HIMSS is to improve healthcare throughout the world. The health care systems’ issues are attackers trying to change the health data, which leads to severe damage in the healthcare system, severe attacks like the ransom ware attack, and lack of cyber security [3]. Additionally, these therapeutic data were collected and funneled from abandons built to store the file folders associated with these documents [4]. Things began to change with the introduction of information technology, and records were transferred to digital media for storage and retrieval, resulting in the creation of electronic healthcare record systems [5, 6]. Automated record management enables easy access to patient records connected to monitoring devices to capture and analyze patient information in the EHR [7]. Quickly, as the EHRs grew in popularity, they began to be used in production and in massive quantities to keep medical records proven to help conduct epidemiological research [7].

The following is a list of EHR requirements:

(a) Interoperability: Interoperability in electronic health records is defined as the capacity of devices and systems to exchange and convert data [8, 9].

(b) Privacy and security: Through the granting of authorizations, security and discretion in the health care setting are meant to empower individuals to control their medical data [10, 11].

(c) Privacy: Privacy is distinct from discretion in that it refers to the measurement of consistent statements or agreements between workers and affected persons. Confidentiality of persistent data is necessary [11].

(d) Admission control: Therefore, remedial records should be restricted to authorized health care professionals and affected persons alone. Affected should control their information and have some say over who has contact with it.

(e) Information sharing: Conversation of remedial archives is necessary when a patient’s treatment is spread among several health care providers; as a result, information is public with other corrective organizations and the administration.

(f) Information Truthfulness and Accessibility: Maintaining data truthfulness entails ensuring its efficacy and consistency. It results in the fact that data is not harmed by unauthorized usage in electronic health records.

1.2 COVID-19

It was discovered in December 2019 in Wuhan, China, and has since banquet to other parts of the country, resulting in a continuing epidemic. The first verified occurrence occurred on November 17, 2019. As of June 18, 2020, over 8.36 million affected cases have been recorded in 188 nations and regions, leading to over 449,000 fatalities. Over 4.09 million individuals have been rehabilitated. Fever, coughing, exhaustion, the littleness of breathe, and harm of smell and taste are all common symptoms. While most cases resolve without complications, some proceed to acute respiratory distress syndrome (ARDS), which may be activated by cytokine storm, [7] multi-organ letdown, septic surprise, or blood clots. The time between acquaintance and development of indications is usually five days but may vary between two and fourteen days.

Individuals may also get infected less often by contracting a polluted superficial and subsequently poignant expression. It is most transmissible during the initial three days following symptoms, although transmission may occur before the onset of symptoms and from those who do not exhibit signs.

1.3 Mode of Spreading

According to 2019 research published in the Journal of the American Medical Association, A giotensin converting enzyme 2 (GCE2) is a membrane exopeptidase found in the coronavirus receptor used to infect human cells (Fig. 1).

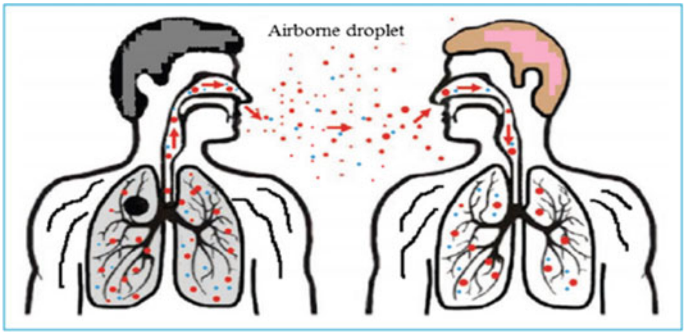


Fig. 1 Transmission of coronavirus via airborne droplets

1.4 Types Coronaviruses

Coronaviruses are a subfamily of the Coronaviridae family. The severity of the illness caused by different kinds of human coronaviruses varies, as does their ability to propagate. Currently, physicians identify seven distinct coronavirus strains that are capable of infecting humans.

1.5 Signs and Symptoms

Although the temperature is the most recurrent symptom of COVID-19, its intensity and exhibition are highly varied, with some elderly, immunocompromised, or severely sick individuals experiencing no fever at all. According to one research, only 44% of patients had a fever before admission to the hospital, but 89% developed fever throughout their stay. The absence of fever does not imply that an individual is disease-free. Cough, lack of taste, tiredness, the littleness of breath, mucus production, strength, and joint aches are other frequent symptoms. Nausea, vomiting, and diarrhea have all been reported at different rates. Sneezing, runny nose, sore throat, and skin sores are less often seen symptoms. In other instances in China, just chest pressure and palpitations were initially seen. It is possible to have a diminished sense of smell or taste abnormalities [10]. In South Korea, 30% of confirmed cases had a loss of smell as a presenting symptom.

1.6 Causes of COVID

COVID-19 feasts mainly when individuals are in adjacent interaction when one individual gasp little precipitations generated by an infected person cough, sneeze, talking, or vocal. The WHO advises a social distance of one meter (3 feet); the CDC in the United States recommends a distance of two meters (6 feet). Individuals may spread the virus even when they exhibit no symptoms, although the frequency with which this occurs is unknown. Asymptomatic infection, according to one estimate, accounts for 40% of individuals affected.

When infected precipitations fall to grounds or surfaces, they may stay communicable if individuals contact polluted surfaces with unwashed hands and then touch their eyes, nose, or mouth. The quantity of active viruses on surfaces diminishes over time, pending it is no lengthier infectious, and exteriors are not considered the primary mode of viral transmission. Although the number of viruses needed to infect through this technique is unclear, it may be identified on copper for up to four hours, cardboard for one day, pliable (polypropylene), and stainless steel for three days (AISI 304). Exteriors may be readily disinfected using home decontaminators that terminate viruses both on and off the anthropoid body. Decontaminators and bleach are ineffective treatments for COVID-19 and may create health issues if not handled correctly, such as within the human body.

Sputum and saliva both contain a high number of viruses. Although COVID-19 is not sexually transmitted, it is believed to be spread through kissing, personal touch, and fecal–oral pathways. Certain medical operations generate aerosols, which facilitate the spread of the infection.

COVID-19 is a novel illness, and the exact mechanisms by which it spreads remain unknown. It is readily transmitted between people—more effortlessly than flu, but not as effortlessly as rubeola. The number of individuals infected with COVID-19 (the R0) by a single person has been extensively estimated. Although the WHO first estimated the R_0 to be between 1.4 and 2.5 (on average, 1.95), a more current assessment determined that the basic R_0 (without control events) is 3.28, and the median R_0 is 2.79. Although the virus may be present in breast milk, it is unclear whether it is contagious or transmissible to the infant.

2 Blockchain

Blockchain and intelligent ledger research have grown in prominence in recent years due to the rise of cryptocurrencies such as Bitcoin and Ethereum. Blockchain technology securely provisions and distributes information in a decentralized, trustworthy, and irreversible way, eliminating the need for intermediaries and eliminating the need for centralized auditing of communications [12, 13]. Transparency in blockchain enables a more straightforward way of accessing ledger-based communications across systems; it links to the diverse processing capacity of various bulges in the blockchain system, making it very influential in terms of computation speed [14]. Blockchain technology encompasses a variety of methods and facilities, including the consensus procedure, hash cryptography, absolute ledger, dispersed peer-to-peer interacting, and withdrawal, which will be discussed in more depth below:

- Consensus procedure: Certain individuals in a blockchain system have unique entree privileges to allowance updated communications in the structure, referred to as the consensus procedure.
- Hash cryptography: To add transactions, a blockchain employs the SHA256 hash algorithm. This is a 64-character extended code created by the NSA. Hash procedures must comprise characteristics such as one-way cryptography, determinism, quicker computing, and the avalanche effect, as well as the ability to resist collisions;
- Immutable ledger: Within a blockchain system, all transactions are noted, and the shared catalog cannot be changed or interfered with;
- Distributed P2P network: To circulate and appraise the information, all communications are transmitted over the system to various users and
- Mining: Miners create hash values for the network by combining blocks of nonce values. To accomplish and receive the reward, this needs a high rate of computing.

The adoption of blockchain knowledge in the health care sector is staged. The first step involves direct connectivity between healthcare providers and the blockchain; all scientific information is recorded and deposited in current well-being information technology schemes. Numerous patient-related data is sent to the blockchain network through API utilizing patient IDs. The inbound transactions are then executed using an intelligent contract inside the blockchain technology. All communications are recorded on the blockchain system using anonymous patient community IDs. The blocks are generated and connected through the immutable ledger. Following that all transactions are committed and assigned a unique identifier. As a result, reverse mining and query processing begin with the well-being provider's APIs. Only non-identifiable affected person information, such as gender, age, and diseases, are kept in the block record. Practical knowledge is inspected to identify novel intuitions.

Use of blockchain to secure HER during COVID-19 Pandemic:

The following sections address and analyze the possibility of using the blockchain scheme to control and relieve the COVID-19 situation.

2.1 Management of Clinical Trial Data

Clinical trial data management is fraught with difficulties due to how companies gather, retain, and exchange data. The blockchain provides a solution to these issues and can fundamentally alter how clinical trials are conducted in the future. Clinical trials must retain data according to laws that require records to be accessible to stakeholders, secure, and immutable. Clinicians and physicians will be able to capture and make accessible real-time health data using blockchain technology. It increases information accuracy, simplifies data sharing, guarantees acquiescence, and provides an audit stream for enhanced confidentiality and information security. Blockchain technology can address various issues by establishing a shared permissioned record of participants, papers, responsibilities, and clinical trial data. It can promote more openness and trust among clinical trial stakeholders. Clinical trials include sharing data between many stakeholders, and building trust between these parties will be critical. For the pharmaceutical sector, there is a need to develop simple models that can educate stakeholders about the blockchain's operation and apply these models to scenarios that may provide economic value with the appropriate partnerships [10].

2.2 Necessary Medicine and COVID-19 Vaccine Supply Chain

The health supply chain may be managed effectively using blockchain technology, especially in pandemic circumstances involving large worldwide cross-border transactions. Until the approved form of the vaccine is ready for sale and marketing, there may be inconsistency in its distribution. Corrupt methods such as manufacturing bogus vaccinations, overcharging, stock accumulation, and so on may be possible. These issues may be resolved effectively by using a blockchain-based medical supply chain. It would demonstrate the provenance of goods and provide increased security via the usage of blockchain.

2.3 Tracing of Contact

Administrations and well-being care institutions are involved in affected person contact monitoring schemes; however, the data gathered may be skewed. Contact tracking has been extensively utilized to help manage the spread of Coronavirus-2019 (COVID-19). It allows the identification, evaluation, and management of COVID-19-exposed individuals, thus limiting the virus's future transmission. Most contact tracing methods, tools, and solutions available today fall short of delivering decentralized, transparent, traceable, irreversible, auditable, secure, and trustworthy characteristics. Data will be consistent and reliable as a result of the usage of blockchains. Blockchain networks allow for the tracking of patient behavior and the provision of real-time information to impacted regions. Additionally, the data may produce a report on the sick and possibly infected population depending on contacts.

2.4 Aggregation of Data

To protect patient privacy and offer more personalized healthcare services, the capacity of blockchain to validate and preserve permanent, real-time data guarantees data integrity. Utilizing a blockchain network enables collecting, storing, and analyzing data on viral spread and control.

2.5 Privacy of User Data

Representatives and health care physicians must collect patient information via patent monitoring and other efforts to aid decision-making while also discussing patient confidentiality and confidentiality concerns. To restore confidence in the system during these troubling days, a delicate balance between highest management and operator confidentiality management must be struck. Blockchain technology may collect and exhibit patient data, monitor affected person treatments, and create public separation degrees while maintaining patient confidentiality.

2.6 Early Discovery of a Susceptible Population

Numerous AI-based triage systems may help patients cope with anxiety. The online bot will assist in identifying early warning symptoms and then direct them toward preventative measures like social withdrawal, hand cleanliness, and so on. Notify users when their symptoms worsen. Confidentiality and privacy of patient information are critical to their personal and societal values being protected. These security and privacy concerns may be successfully addressed via blockchain-based construction.

3 Summary and Discussion

The electronic health record (EHR) is a computerized archive of a patient's medical history. It has resolved a slew of problems relating to the management and security of data. This article discussed different EHR standards, EHR system problems, and the potential of blockchain to address these concerns. We then looked at the various application situations for blockchain-based EHRs in epidemic supervision, including contact outlining, information aggregation, information sharing, user information privacy, COVID certificates for diseased and improved individuals, forecast of contagion people progression, and efficient and reliable supply chain supervision for vaccines and other indispensable supplies. The integration of blockchain knowledge and machine learning systems enables developing a generalizable prediction system that may aid in the containment of pandemic risk on national territory.

4 Conclusion

COVID-19 initiates a worldwide pandemic of infection, the most devastating in human history. Without specific medications or vaccinations, identifying and isolating the source of infection is the most excellent option for slowing the spread of the virus and lowering infection and mortality rates among the population. The procedure of tracking the virus encounters three major roadblocks: (1) The electronic well-being record of each affected person is kept in a centralized database that may be stolen and the infection data changed with (2) A third party or organization may get access to the affected user's sensitive personal information (3) Existing infection tracing methods do not allow for the tracking of illnesses in many dimensions. The system supports either location-based tracing or individual-based tracing. To safeguard users' privacy, they may utilize the blockchain network to anonymously transmit their visit data and health condition. By enabling dependable, accurate, and secure data storage and sharing, blockchain technology may effectively address the limitations of conventional EHR systems and contribute to the efficient and effective management of the COVID-19 epidemic.

References

1. Valerie S, Prater (2014) Confidentiality, privacy, and security of health information: Balancing interests—Health Informatics Online Masters | Nursing and Medical Degrees
2. Palvia P, Jacks T, Brown W (2015) Critical issues in EHR implementation: provider and vendor perspectives. Commun Assoc Inf Syst 36:707–725. https://doi.org/10.17705/1CAIS.03636
3. Marquez G (2017) The history of electronic health records | Elation Health | Clinical First Electronic Health Record | Elation Health. Elation Health
4. Gordon WJ, Catalini C (2018) Blockchain technology for healthcare: facilitating the transition to patient-driven interoperability. Comput Struct Biotechnol J 16:224–230. https://doi.org/10.1016/j.csbj.2018.06.003
5. Adel E, El-Sappagh S, Barakat S, Elmogy M (2018) Distributed electronic health record based on semantic interoperability using fuzzy ontology: a survey. Int J Comput Appl 40(4):223–241. https://doi.org/10.1080/1206212X.2017.1418237
6. Alam S, Shuaib M, Samad A (2019) A collaborative study of intrusion detection and prevention techniques in cloud computing. In: Lecture notes in networks and systems, vol 55, pp 231–240
7. Shrivastava U, Song J, Han B (2019) The implications of patient data security considerations for HER interoperability and downtime recovery
8. Barrick G (2019) 4 reasons why EHR interoperability is a mess (and How to Fix It) | Datica. Datica
9. Hardin T, Kotz D (2019) Blockchain in health data systems: a survey. In: 2019 6th international conference on internet of things: systems, management and security (IOTSMS), pp 490–497. https://doi.org/10.1109/IOTSMS48152.2019.8939174
10. Siddiqui ST, Alam S, Shuaib M (2019) Cloud computing security using blockchain. J Emerg Technol Innov Res 6. [Online]. Available www.jetir.org
11. Kathole AB, Halgaonkar PS, Nikhade A (2019) Machine learning and its classification techniques. Int J Innov Technol Explor Eng (IJITEE) 8(9S3). ISSN 2278-3075
12. Kruse CS, Smith B, Vanderlinden H, Nealand A (2017) Security techniques for the electronic health records. J Med Syst 41(8). https://doi.org/10.1007/s10916-017-0778-4
13. Samad A, Shuaib M, Rizwan Beg M (2017) Monitoring of military base station using flooding and ACO technique: an efficient approach. Int J Comput Netw Inf Sec 9(12):36–44. https://doi.org/10.5815/ijcnis.2017.12.05
14. Siddiqui ST, Shuaib M, Bokhari MU (2019) Web-based requirements management tools for software development: a study. In: Proceedings of 12th INDIACom-2018, February 2019. IEEE, pp 10–15
15. Bhartiya S, Mehrotra D, Girdhar A (2016) Issues in achieving complete interoperability while sharing electronic health records. Elsevier. https://doi.org/10.1016/j.procs.2016.02.033
16. Catalini C, Gans JS (2016) Some simple economics of the blockchain. SSRN Electron J. https://doi.org/10.2139/ssrn.2874598

17. Evans RS (2016) Electronic health records: then, now, and in the future. Yearb Med Inform (Suppl 1):S48–S61. https://doi.org/10.15265/IYS-2016-s006
18. O'Dowd (2017) Lack of EHR interoperability standards challenges health IT

Chapter 18
Secure Hybrid Approach for Sharing Data Securely in VANET

Atul Kathole and Dinesh Chaudhari

1 Introduction

A network is a collection of two or more computers that are connected to share resources. There are two distinct kinds of networks. The first is a network that is connected through wires. The other kind of network is wireless. The following subcategories apply to wireless networks: One kind of network is infrastructure based, dubbed cellular. The last kind of network lacks infrastructure, dubbed VANET. There is no need to safeguard users' mobility behaviors or movement patterns in wired networks since equipment such as PCs are always static and do not move from one location to another. However, sensitive data should be protected from attackers in wireless settings. Otherwise, an opponent may create profiles of people based on their actions and exploit that knowledge to endanger or damage individuals. Finally, the difficult challenge with ad hoc networks is maintaining privacy while using low-power wireless devices and low-bandwidth network connections.

The term "privacy-preservation" refers to the practice of safely storing private data. To eavesdrop on conversations via wired networks, one must first obtain access to the wire's connections. The attacker needs a suitable transceiver to intercept the receiver's wireless signal without being noticed. Anonymity, invisibility, and unlikability are all privacy-preserving characteristics. Anonymity is the condition of not being individually identifiable within a group of topics known as the anonymity set. Unlikability between two or more things of interest implies that from the attacker's viewpoint, these items of interest are no more or less linked after his observation than before to his observation inside the system containing these and potentially additional items. Unobservability refers to the inability of items of interest (IOI) to

A. Kathole (✉)
Department of Computer Engineering, PCCOE, Pune, India

D. Chaudhari
Department of Computer Science and Engineering, PCCOE, Pune, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_18

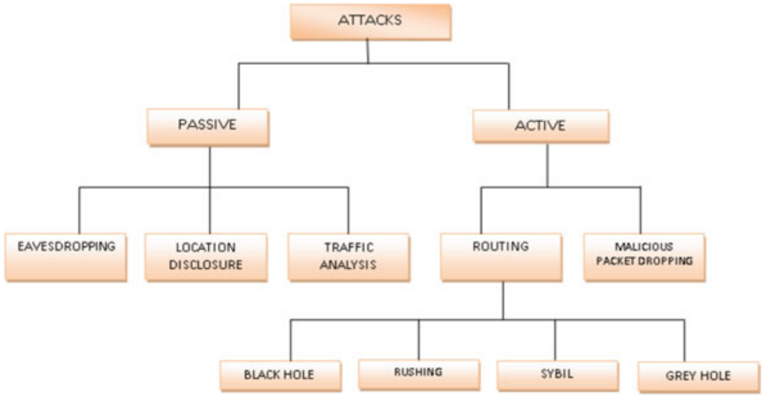


Fig. 1 Classification of attacks

be distinguished clearly from any other IOI of the same kind. Strategies often fail to comprehend the true nature of an unfavorable case's reasons. This leads to a high number of false positives for nodes that are not malicious and low detection rates for malicious nodes. Such vulnerabilities exist as a result of the assumptions made by these confidence-based security methods. After all, packet losses occur only due to malevolent behavior on the part of misbehaving nodes. These are, in any case, due to a variety of factors, including insufficient agility propagation, congestion, and wireless connections. Without a fine-grained examination of packet losses, conventional detection methods may result in erroneous confidence estimations, mainly when mobile nodes and data rates are high [1].

This section categorizes assaults based on their perceived characteristics (Fig. 1).

The rest of the research is organized as follows: Sect. 4 provides nodes for harmful identification systems in II, Related Work, and III Study Process. V details the simulation's effects. Finally, VI contains a summary of the study.

2 Research Method

We recommended the hybrid bait detection system, or HBDS, as a harmful detection tool in this post. This is just a parameter for negative identification nodes to resolve the problems encountered by earlier detection methods, which were mainly dependent on packet losses. The HBDS presented is a two-stage detection method in which the MANET defense provides and reduces the nodes that are malicious detection mistakes. Numerous security mechanisms make it impossible to identify rogue nodes with certainty [2].

The graph illustrates the node mobility about the packet delivery ratio, throughput, and end-to-end latency. In the presence of Sybil attack, HBDS has a packet delivery ratio of 96.5 percent, a throughput of 38.37 percent, and an end-to-end latency of

0.34 percent, which is comparable to FGA. Utilizing the suggested method enables the network's overall preference to be improved to achieve maximum throughput in the shortest amount of time [3].

3 Malicious Activity Detection

The primary goal of this study is to offer insight into the malicious node detection mechanism used in MANETs to enhance security and performance. The software is built on the HBDS framework, a capable of defending against a variety of MANET assaults. To optimize end-to-end latency and PDR performance, the hybrid CBDS (HBDS) technique is needed. The proposed work versus existing HBDS methods would significantly improve performance against a range of network attacks [4].

We utilize the suggested method to avoid a Sybil attack on a malicious node inside a particular MANET.

Algorithm

The value assigned to a node by a source node will be computed and compared to the actual behavior of each node in order to determine the correctness of the underlying trust-based framework that has evaluated credibility criteria. The pseudocode for our HBDS technique is presented after the algorithm [3].

The above algorithm will consider the different parameters with packet loss and PDR to evaluate the node working. If the particular node drops the packet gather, then the set threshold value and PDR are also greater than 0.5 that time node id is captured and stored in a separate table as it can be a malicious node on an above predication [6].

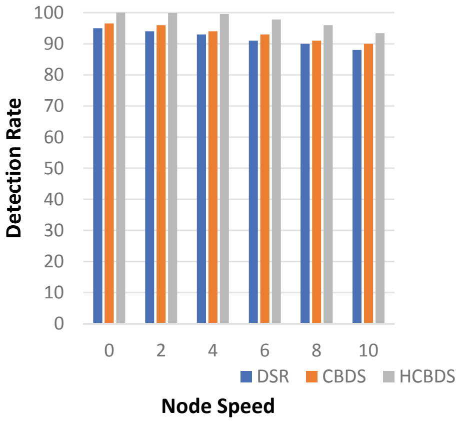
4 Simulation and Results

Below, you will find some analysis of HBDS performance using corresponding methods.

(1) **Detection Rate**

Compared to the other plan's rate, our HBDS framework offers a higher level of identification efficacy. Similarly, Fig. 2 illustrates the recognition rate as a function of expanding hub speed for the HBDS methodology and other techniques. In contrast, Fig. 2 illustrates the location rate as a function of expanding hub thickness. The amount of information associations inside an organization increases as the hub thickness increases, as more bundles are lost due to effects. The alternative paradigm views these parcel drops as upstream actions from real hubs. Subsequently, as shown in the figure, the identification rate is more significant in our HBDS cycle than in the other cycles [7].

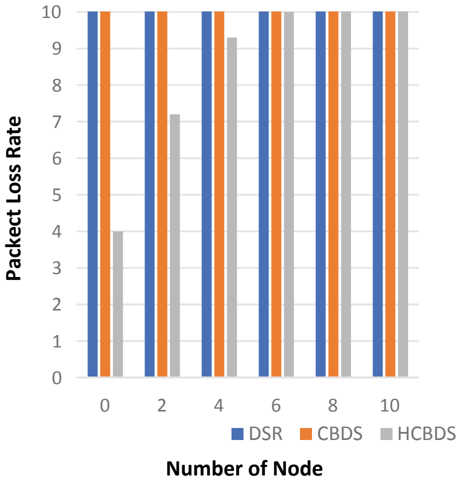
Fig. 2 Effect of a node moving speed and density on detection rate. (a) Detection rate versus node speed



(2) Packet Loss Rate

The packet loss rate as a function of increasing node speed is shown in Fig. 3 for the HBDS schemes, respectively. Our HBDS system, as shown in the picture, has a lower packet loss rate than the FGA system. In the HBDS system, more reliable nodes are often chosen for routing, resulting in fewer error packets and a higher package transmission ratio.

Fig. 3 Effect of node moving packet loss rate



5 Conclusion

While there are many ways for defending ad hoc networks against such attacks, traditional preventative measures have significant limitations and disadvantages in this area. Numerous conventional methods are imprecise in their use. DSR often fails to remove rogue nodes during the route discovery process, failing Sybil attacks to deliver all data packets to the target.

In conclusion, we can prevent Sybil attacks through our suggested methods, and the observed increase in throughput and packet delivery ratio is noteworthy.

References

1. Kathole AB, Chaudhari DN (2019) Pros and cons of machine learning and security methods. 21(4). http://gujaratresearchsociety.in/index.php/JGRS. ISSN 0374-8588
2. Kathole AB, Halgaonkar PS, Nikhade A (2019) Machine learning and its classification techniques. Int J Innov Technol Explor Eng (IJITEE) 8(9S3). ISSN 2278-3075
3. Kathole AB, Chaudhari DN (2019) Fuel analysis and distance prediction using machine learning. Int J Future Revol Comput Sci Commun Eng 5(6)
4. Hasrouny H, Samhat AE, Bassil C, Laouiti A (2017) VANet security challenges and solutions: a survey. Veh Commun 7:7–20
5. Yaqoob I, Ahmad I, Ahmed E, Gani A, Imran M, Guizani N (2017) Overcoming the key challenges to establishing vehicular communication: is SDN the answer. IEEE Commun Mag 55(7):128–134
6. Ahmad I, Noor RM, Ali I, Imran M, Vasilakos A (2017) Characterizing the role of vehicular cloud computing in road traffic management. Int J Distrib Sens Netw 13(5). Art. no. 1550147717708728
7. Khan MS, Midi D, Khan MI, Bertino E (2017) Fine-grained analysis of packet loss in MANETs. 2169-3536 2017 IEEE
8. Ahmad I, Ashraf U, Ghafoor A (2016) A comparative QoS survey of mobile ad hoc network routing protocols. J Chin Inst Eng 39(5):585–592

Part IV
Machine Intelligence

Chapter 19
MetaEfficientNet: A Few-Shot Learning Approach for Lung Disease Classification

Shravani Nimbolkar, Anuradha Thakare, Subhradeep Mitra, Omkar Biranje, and Anant Sutar

1 Introduction

Lungs play a major role in the functioning of the body. It becomes very much necessary to ensure their proper functioning. Lung diseases are a big problem in today's society, where pollutants and dangerous viruses play a big part. The manual diagnostics of these kinds of diseases, although possible, has some drawbacks, and thus we think it is the need of the hour to implement automated systems, using the latest available techniques. Lung diseases also known as respiratory diseases registers as one of the leading causes of mortality today. Around 65 million people suffer from one of the other forms of lung disease and about 3 million die per year [1]. The severity of chronic respiratory diseases can be understood by considering the very latest example of COVID-19. There are different types of lung diseases and diagnosing them using CXR with the naked eye by a specialist or by a naive person may give us misleading results. Pneumonia and Tuberculosis may look similar to an untrained person but require completely different diagnosis and treatment. Early diagnosis, which is critical in the treatment of any disease, improves doctor-patient communication, enables for more focused diagnosis and therapy, and, most significantly, aids in the detection of early warning symptoms.

S. Nimbolkar · A. Thakare · S. Mitra · O. Biranje · A. Sutar (✉)
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune, India
e-mail: anant.sutar17@pccoepune.org

S. Nimbolkar
e-mail: shravani.nimbolkar17@pccoepune.org

S. Mitra
e-mail: subhradeep.mitra17@pccoepune.org

O. Biranje
e-mail: omkar.biranje17@pccoepune.org

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_19

Medical imaging, such as X-ray scanning, bundled with artificial intelligence (AI) technologies, is an encouraging and logical tool for the detection and classification of lung diseases. Deep learning and machine learning methodologies have proved crucial in identifying such diseases. Methods such as CNN where we are able to diagnose ailments such as Tuberculosis [2], Pneumonia [3], COVID-19 [4], and many others are having human-like abilities to classify such diseases. Further to improve accuracy and save time and effort, the concept of transfer learning was introduced where it was able to classify disease using networks and weights. [5] contributed a large-scale survey of deep learning methods to identify lung diseases. Albeit such promising results, a problem that the above methods face is the quantity of data. Deep learning has conveniently set a new standard for object detection, however, to achieve such benchmarks it needs tons of data to build its intuitive behavior. Also in some domains, the availability of data is scarce. The solution for such problems is meta-learning. Meta-learning refers to adapting and learning quickly through the use of a few examples. One-shot and Few-shot learning use one or more samples to learn and take that information to achieve tasks such as detection and classification. In this paper, we propose a meta-learning approach to classify lung diseases. The following paper is divided into sections. Section 2 gives an account of related studies such as transfer learning and meta-learning techniques. Section 3 provides details on the proposed model. Section 4 gives the algorithm of the proposed mode. Section 5 gives evaluation parameters used. Section 6 discusses the result of our experimental studies and a comparative analysis.

2 Related Study

2.1 Transfer Learning

Transfer learning is an approach in deep learning wherein the pre-trained models are used as the first step in computer vision. It utilizes the information gathered from one task to solve tasks related to it. Pre-trained models are trained rigorously on a large image dataset like ImageNet dataset with 1.2-million color images and 1000 classes, which is called a base dataset consisting of a wide variety of samples. The knowledge acquired throughout the process is then transferred while performing new tasks. Since developing neural network models on complex problems require high computations and time, transfer learning reduces the time it takes to construct and train a model by reusing modules of previously trained models. We are able to import transfer learning models and apply them to problems such as image classification, text classification, etc. Over the years, the pre-trained model approach has gained traction due to its ease and efficiency. First, we select a source model from all the available models. There are numerous models present such as VGG16, ResNet, DenseNet. Once we select our source model we then according to our convenience can modify or tune it. These features allow people to construct models that suit their needs and

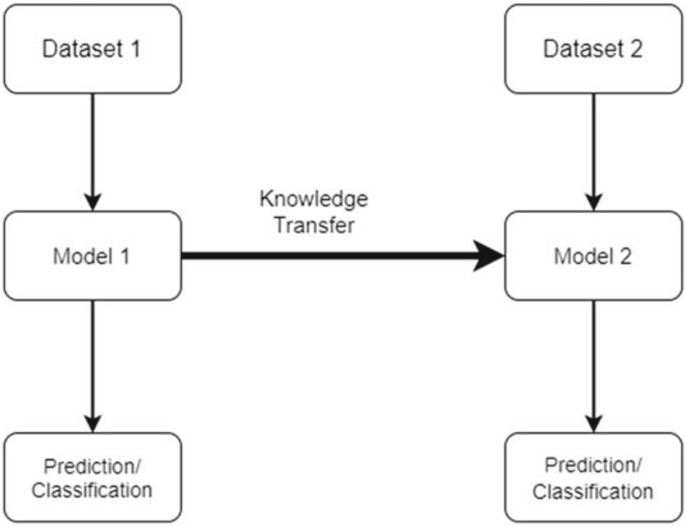


Fig. 1 Transfer learning

make such models flexible. Then we can use the tuned model to solve problems such as detection, classification or optimization (Fig. 1).

The following study was conducted to bolster the importance of transfer learning. Hashmi et al. [6], proposed a high-performance model applicable to digital chest x-ray images. It introduced a weighted classifier approach, which combined weighted predictions from ResNet18, InceptionV3, DenseNet121, etc. in an optimal way. To fine-tune the models, transfer learning was used. The Guangzhou Women and Children's Medical Center pneumonia dataset was used in this study. To increase the training dataset in a balanced way, partial data augmentation was employed. The model achieved a test accuracy of 98.43% and an AUC score of 99.76. A major drawback of this approach was the scarcity of available resources resulting in overfitting. And it adversely affected models' generalization ability. The diagnosis of the disease requires a deep understanding of the radiological features from chest x-rays. Militante et al. [7], proposed a methodology that employs efficient approaches of 6 CNN models, to recognize and predict patients affected and unaffected with the disease, using a chest x-ray image. The models used were: LeNet, VGG16, StridedNet, GoogLeNet, AlexNet, and ResNet-50. The Radiological Society of North America (RSNA) dataset, which consisted of a total of 28,000 chest x-ray images, was used. The images were in JPEG format, having maximum dimensions of 1024 × 1024. All the images were categorized into two classes: infected and not infected. All the models attained a training accuracy of 96–98%, with GoogLeNet and LeNet obtaining the highest, i.e., 98% accuracy for performance training. This study can also consider the optimization of hyper-parameters to improve the accuracy of the model. Ahsan et al. [8], proposed applying the VGG model on the chest x-ray dataset to spot if a patient is suffering from TB. The dataset used by the author consisted of

1324 CXRs and text data from Shenzhen and 276 from Montgomery datasets. The dataset was divided into training and testing datasets with a 75 to 25 ratio, respectively. The VGG16 model was used however instead of using a soft-max layer the author preferred to use a sigmoid layer as they only had two outputs. The model was able to attain 80% accuracy without augmentation and later achieved 81.2% with it. TFLearns data augmentation of TensorFlow was used for augmentation which helped in the making of new images and also alleviates data overfitting in the model. The author mentioned that the low accuracy was due to partial augmentation of the dataset and it could be improved by running the model on a high configuration system.

2.2 Meta-Learning

Meta-learning is a subfield of machine learning that focuses on classification and regression tasks and employs a learn-to-learn method. Because the algorithm learns from the output of other algorithms, it is also known as a meta learner. A successful meta-learning model should be able to adapt or generalize to new tasks that have not been encountered yet during the training phase. There are three types in meta-learning: learning in metric space, learning the initializers, and learning the optimizers [9].

Learning in metric space. This method is similar to that used in closest neighbor algorithms. The model learns to find an optimal similarity distance function. It employs the kernel density estimation concept, in which the probabilities predicted for a particular input are determined as a weighted sum of its labels, with the weights generated by a kernel function. This approach attempts to determine the degree of similarity between a given image and the image in the support set, which are samples from the database. The few-shot learning is a very popular problem that is solved with metric-based learning.

Learning the initializers. The weights of a neural network are randomly initialized at the beginning of the training phase. These weights are changed throughout training by reducing losses via gradient descent. However, the convergence can occur early if the weights are initialized to their optimal values or close optimal values. We can find these initial optimal values for weights using algorithms like Reptile, Meta-SGD, and MAML.

Learning the optimizers. In this method, we train the optimizers. The neural networks are further optimized after the training phase by lowering losses and training on larger datasets. However, because few-shot or one-shot learning models are trained on a significantly smaller collection of data, gradient descent may fail. In optimization-based meta-learning, we can have the base model for learning and a meta-model for optimizing the base model.

2.3 One-Shot and Few-Shot

One-shot and few-shot models, like the human brain, can learn to differentiate between images belonging to multiple classes. With metric-based meta-learning, few-shot learning is widely employed. Few-shot learning comes in four flavors: zero-shot learning, one-shot learning, and N-shot learning. The N-way-K shot problem is classified as a typical few-shot learning problem, where the N refers to the number of class labels and K refers to the number of samples provided for every class. For the N-way-K shot problem, the task which includes N-way K samples is known as the support set. Few more examples of the same classes are included in the Query set which is used for performance evaluation of the task. One-shot can also be termed as N-way one-shot problem where the model is fed with a single sample for each class. Siamese neural networks are the most commonly used metric-based one-shot and few-shot learning algorithms as similarity networks. It consists of a pair of convolutional neural networks and outputs the similarity measures for given samples. (Shuti Jadon, 2020) has presented a detailed study of deep learning in few-shot learning [10].

2.4 Siamese Neural Network

Siamese networks are a class of neural networks and fall under the category of meta-learning or learning to learn. The network is a combination of the same embedding models which produce an embedding for each input image, i.e., each image is fed to one embedding model. The produced embeddings for each image is compared with embeddings of others as per the dissimilarity measure chosen. This dissimilarity measure is then passed to a contrastive loss function, which then updates the weights of the embedding model (Fig. 2).

Li et al. [11] discusses the schematics of Siamese networks on medical images taking the Euclidean distance between the embedding outputs produced by two images as a measure of their dissimilarity. This dissimilarity is used to train the embedding model to produce better embedding so that similar images have the least dissimilarity, and images from different classes will have higher dissimilarity. For this they used the ROP dataset which contained x-ray images of knee joints of patients with pain severity (0–5) of the area as the class or category of the x-ray. They achieved an AUC of 0.81 (95% CI 0.77–0.84) with Cohen's kappa 0.64 (95% CI 0.58–0.68) and an AUC of 0.90 (95% CI 0.88–0.92) with Cohen's kappa 0.66 (95% CI 0.61–0.71) using the Euclidean distance difference. The network trained for knee osteoarthritis achieved an AUC of 0.90 (95% CI 0.86–0.83) and 0.88 (95% CI 0.84–0.92) with Cohen's kappa 0.47 (95% CI 0.40–0.54) and 0.41 (95% CI 0.35–0.48), respectively, using the Euclidean distance difference. For ROP, the conventional neural network achieved a linear Kappa of 0.61 (95% CI 0.55–0.66). For knee osteoarthritis, the conventional neural network achieved a linear Kappa of 0.46 (95% CI 0.39–0.54).

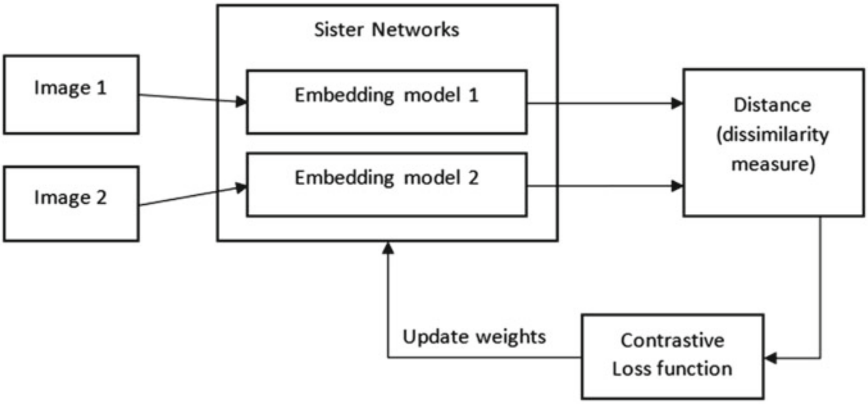


Fig. 2 General structure of a simple Siamese network for meta-learning

It found overlap in 95% in the confidence intervals of the Siamese network, thus stating the performance of Siamese networks on binary change detection.

Mohammad Shorfuzzaman et al. [12] pointed out the high data required for training the traditional deep learning algorithms, and if the dataset is insufficient, then it needs to be handled by using data augmentation techniques like GANs. The newly generated data if hand-tuned then has a high chance of overfitting, or if generated through GANs face challenges in simulating real patient data which leads to unanticipated bias during model testing as the augmented data may be impossible to exist in nature. They used both loss functions, binary cross-entropy, and contrastive loss function. The chest x-ray images were rescaled to 100 × 100 and each pixel value was converted from [0, 255] to [0, 1]. The model was evaluated on different metrics and got the highest accuracy of 95.6% for the 3-way 10-shot learning approach using contrastive loss and 93.8% for the 3-way 10-shot learning approach using cross-entropy loss. The model was compared with other pre-existing models like InceptionV3, Xception, InceptionResNetV2, and VGG16, the proposed model outperformed all of the mentioned models. The model was also evaluated on a 2-way 10-shot approach and got an accuracy of 96.5%.

Prayogo et al. [13] discusses the complex nature of the existing methods for the diagnosis of pneumonia disease. Most of the studies on image classification using deep learning approaches use CNN. CNN has been used in several problems like classifying stroke, type of muscle, and abdominal ultrasound images. They proposed a model for the classification of chest x-ray images into Normal, Bacterial Pneumonia, and Viral Pneumonia. The dataset used in this study was Labeled Optical Coherence Tomography (OCT) and chest x-ray images, which contained 5863 images. They used a convolutional model as the embedding model and followed by a fully connected network that accepts inputs from the embedding model and outputs the distance measure. For training the weights the cosine distance was used as the dissimilarity function. The model resulted in an accuracy of 80.03% for the 20-shots approach.

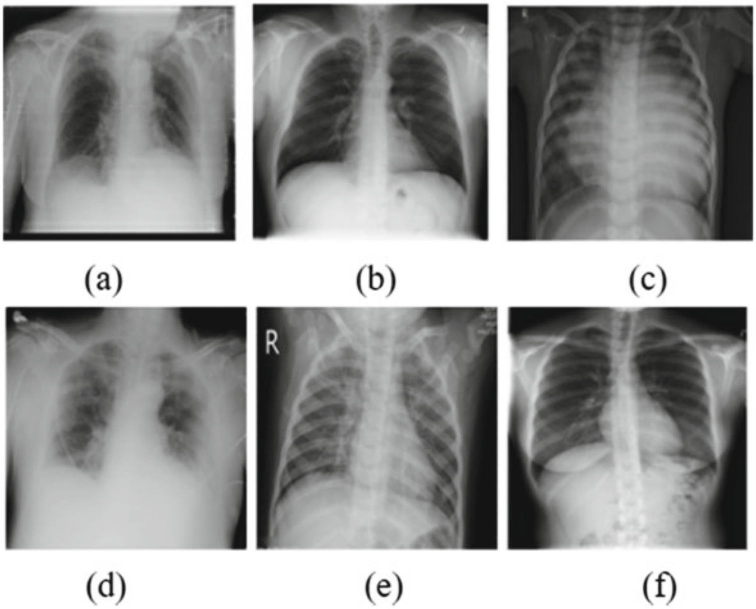


Fig. 3 Illustrates some sample images from the COVID-19 radiography and pneumonia dataset **a**, **d** COVID-19. **b**, **f** Normal. **c**, **e** Viral Pneumonia

3 Proposed System

3.1 Data Acquisition

For this study, we used the COVID-19 Radiography dataset and chest x-ray images dataset [14, 15] which contains chest x-rays images categorized into three classes: Normal, COVID-19, and Viral Pneumonia containing a total of more than 15,000 images (Fig. 3).

3.2 Siamese Neural Network Using Pre-Trained Models

In general, we use supervised metric-based learning to train the Siamese net to learn the image representation in feature space, and then use it for few-shot learning without further training. The Siamese network is made up of two parallel pre-trained CNN embedding models, also known as sister networks, in our proposed architecture. The "twin" here signifies that these two networks have the same parameters and weights. Each network creates an embedding vector of size 128 of its individual inputs after being given a pair of images. These sister networks are then trained to

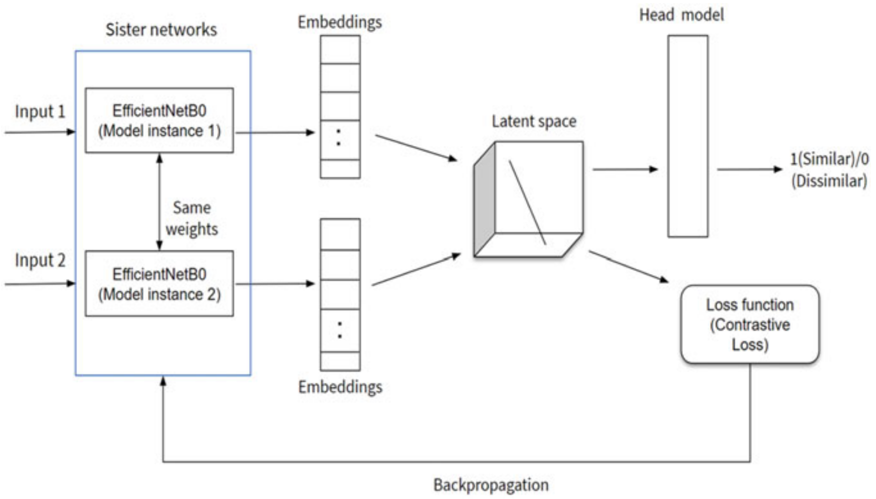

Sister networks
Embeddings
Head model
Input 1
EfficientNetB0 (Model instance 1)
Latent space
1(Similar)/0 (Dissimilar)
Same weights
Input 2
EfficientNetB0 (Model instance 2)
Loss function (Contrastive Loss)
Embeddings
Backpropagation

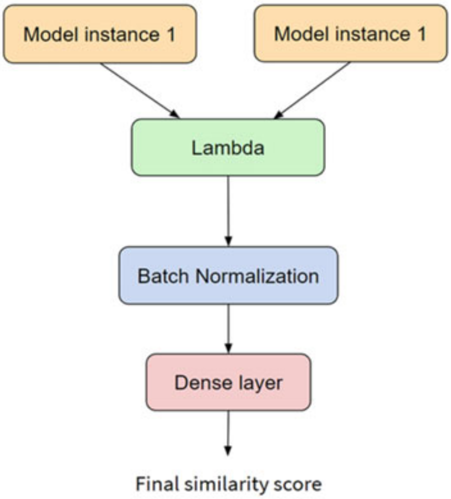
Fig. 4 The proposed architecture, Siamese network using EfficietNetB0 architecture. The blue box envelopes the sister network which is fed with a pair of images. The outputs of that network are the feature embedding of two images which are given to the loss function

maximize the distance between input embeddings for different classes while minimizing the distance between embeddings for similar classes. An energy function sits atop the sister network, calculating the metric between the embeddings from these twin networks. In this paper, we have proposed a Siamese neural network which uses EfficientNetB0 architecture in the Siamese neural network as shown in Fig. 4.

Head model. The final result is produced by stacking the head model on top of the two sister networks, as shown in Fig. 5. It has a Lambda layer that takes the embeddings vectors of size 128 from the sister models as input and computes the Euclidean distance between the embeddings. The lambda layer is followed by the batch normalization layer which normalizes the input coming from the lambda layer. The dense layer is the output layer with sigmoid activation function which produces the final similarity score.

Contrastive Loss function. For the proposed model, learning is done using the contrastive loss function. The choice of the loss function is a crucial step because in the training phase the loss function makes the base model learn to obtain the feature embedding of the input image. Various loss functions, such as triplet loss, contrastive loss, and circle loss, are used to train Siamese networks. The contrastive loss function, which we used in our study, requires a pair of images as input. The function penalizes the base model for generating false feature embeddings based on the label given to the pair of input images. In a nutshell, the loss comes out to be low for similar or closer embedding vectors for images of the same class and high for dissimilar embedding vectors for the images of different classes [16]. Mathematically it is given as:

Fig. 5 The head model



$$\text{Contrastive loss} = Y(D)^2 + (1 - Y)\max(\text{margin} - D, 0)^2 \tag{1}$$

In the above equation, Y is the true label which will be 1 when two input images will be the same and 0 if those are dissimilar. D is our Euclidian distance between two embeddings. $D = ||I_1 - I_2||_2$ where I_1 and I_2 are the two feature embeddings of two inputs. The margin is to hold a constraint over the loss; when two input values are dissimilar, and if their distance is greater than a margin, then they do not incur a loss.

Transfer learning and Siamese neural networks. Transfer learning discussed earlier in Sect. 2.1. is an advanced and smart approach as compared to traditional machine learning, wherein the pre-trained models are used for a new but similar task. Transfer learning works better than machine learning if the model features gained from the previous task are generic enough. Transfer learning provides the distinct benefit of speeding up the model training process and fine tuning the results. In our study, we found that transfer learning has a great potential to leverage the performance of Siamese neural networks and few-shot learning. We experimented with the two very famous pre-trained architectures and fine-tuned them to suit our CXR dataset to avoid training the network from scratch.

These two differ not only by their architecture and depth but also by their size. The VGG16 with size 549 MB due to its depth and a large number of fully connected layers is pretty heavier and more complex in its structures than EffiicentNetB0 of size 29 MB. The two architectures are discussed below in detail.

VGG16. The VGG16 is famously used for image classification especially for medical imaging [17]. The VGG16 network is considered one of the most important CNNs for image classification because of its deep yet simple architecture, which gives it robustness against overfitting while providing good performance. Due to the

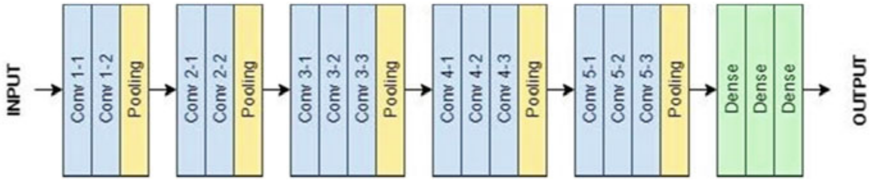


Fig. 6 Standard VGG16 pre-trained model with 16 convolutional layers [20]

smaller kernel size of VGG16, it can extract intricate features in the CXR which are crucial for image classification [18]. The VGG16 has 16 layers out of which 13 are convolutional layers with the Relu activation function and 3 are fully connected layers, with all kernel sizes of 3 × 3. Each convolution layer is followed by a max-pooling layer with all 2 × 2 kernel sizes. Convolution layers function as automatic feature extraction that stores training weights. The next layer is 3 fully connected layers (FC) which are the final layer as a classifier [19]. A detailed architecture of VGG16 is shown in Fig. 6.

EfficienNetB0. EfficientNet was first introduced by Tan and Le [21] in 2019. These models are based on simple and highly effective compound scaling methods. This method enables to scale up a baseline ConvNet to any target resource constraints while maintaining model efficiency, used for transfer learning datasets [22]. EfficientNet has been proposed to improve the performance of CNNs by scaling in three dimensions, i.e., width, depth, and resolution using a set of fixed scaling coefficients that meet some specific constraints [23]. In general, EfficientNet models achieve both higher accuracy and better efficiency over existing CNNs such as AlexNet, ImageNet, GoogleNet, and MobileNetV2. The detailed architecture of EfficientB0 is shown in Fig. 7.

CNN baseline model. Convolutional neural networks (CNNs) were specifically created for working on image data. CNN's tries to extract the features in the image using various filters and kernels. The design is inspired by the visual cortex of the human brain. When an image is fed to the convolutional neural net it processes the image and captures the spatial and temporal dependencies as the filters (Fig. 8).

convolute across the entire image. As a baseline model, we have used a simple CNN, illustrated in Fig. 7, as the embedding models for our Siamese net. The CNN consists of two 2D convolutional layers, two average pooling layers, two batch normalization layers, and one dense layer at the end with tanh activation function.

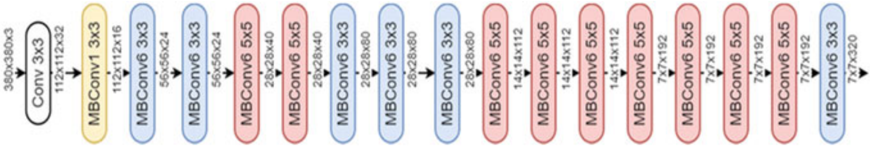


Fig. 7 The architecture of efficientNetB0 [24]

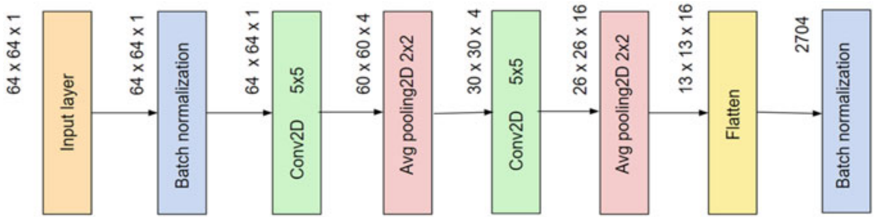


Fig. 8 Architecture of CNN baseline model

4 Methods

We have to classify the CXR images into one of the three classes namely, Normal, COVID-19, and Pneumonia. After initializing model parameter *p* and input data (pairs of images), the model is getting trained for *nepochs* with batch size *N*. In our proposed model, we have used EfficientNetB0-based Siamese net which takes a pair of images as input and gives the embeddings of each image which are denoted as *em1, em2*. After concatenating these embeddings, the Euclidean distance between them is calculated which is denoted as *dist*. The head model which is denoted as *H* receives the *dist* as input and outputs a similarity score given as *predicted_similarity* in the algorithm below. This score together with the true label for the pair of images is given to the Contrastive loss function for calculating the loss *L* and the model parameter is updated with new weights *w*.

4.1 Algorithm for Training

```
Input: Dataset D, Batch size N, Number of epochs nepochs, fine-tuned EfficientB0 model M with parameter p,
Head model H, Loss L, margin m

Initialize: img1, img2, image_pairs, true_similarity_label, predicted_similarity, dist for training

image_pairs, true_similarity_label = get_pairs(D)
$p_0 = w_0$

For i do nepochs
    for b do getBatches()
        img1, img2 = image_pairs
        em1 = M(img1)
        em2 = M(img2)
        $dist_b$ = euclidean((em1,em2))
        predicted_similarity = head_model($dist_b$)
        $L_b$ = ContrastiveLoss(predicted_similarity, true_similarity_label,m)
        Update parameter $p_b$ with new weight, w
    end for
end for
```

4.2 Evaluation Metrics

In the confusion matrix of a multi class classification problem, the columns are for actual class labels and rows are for predicted class labels. From this accuracy matrix, we can calculate different performance measures such as precision, recall, F1-score, and AUC.

Precision. Precision is nothing but the exactness of the classifier, the number of true positive labels divided by total positive labels. In our case for COVID-19 classification the precision would be calculated as true positive for COVID-19 divided by the total positives for COVID-19. Total the formula for precision is given as:

$$\text{Precision} = \frac{n(\text{True Positive labels})}{n(\text{True Positive labels}) + n(\text{False Positive labels})} \tag{2}$$

Recall. The recall is another measure of performance which is also called sensitivity or true positive rate is the measure of classifiers completeness. It can be calculated by dividing the number of true positive labels by the sum of true positive and false negative labels. In our case for COVID-19 classification, the recall would be the true positive cases of COVID-19 divided by the same and false negative cases of COVID 19. The formula for recall is given as:

$$\text{Recall} = \frac{n(\text{True Positive labels})}{n(\text{True Positive labels}) + n(\text{False Negative labels})} \tag{3}$$

F1-Score. F1-score or F-score also called as the harmonic mean of precision and recall is a way to combine precision and recall. F1-score is often reliable when an uneven class distribution seeks a balance between precision and recall especially when there are more labels of true negatives. F1-score for multi class classification is we do not compute overall F1-score, instead we calculate F1-score per class in one vs rest manner.

$$F1\text{ score } = 2 * \frac{\text{Precision}(\text{class} = a) * \text{ Recall}(\text{class} = a)}{\text{Precision}(\text{class} = a) + \text{Recall}(\text{class} = a)} \tag{4}$$

Area Under Curve (AUC). AUC is the area under the ROC curve, is that measure of the flexibility of a classifier to tell apart between classes and is employed as a summary of the ROC curve. If AUC = 1, then the classifier is in a position to discern between all the positive and therefore the negative class points accurately. However if AUC had been 0, then the classifier would not have been able to classify accurately and would have classified all positives as negatives and all negatives as positives.

5 Experimentation and Results

5.1 Preprocessing

The images used in this study were first converted into grayscale format and then resized to 64 × 64 for faster processing of data and faster training of the model. The resized image was then normalized using rescaling by Min–Max scaler (minimum value is 0 and maximum value is 255 for every pixel) so that higher and lower valued pixels will not generate bias in the training of the model. The dataset was found to be imbalanced which was made balanced using sampling methods.

5.2 Training

We evaluated CNN (baseline), MetaEfficientNet, VGG16 and compared their results. We trained all the three models on the benchmark datasets and observed their losses and accuracies throughout the training and testing phase. The standard number of epochs were 8 with a batch size of 16. The train and test ratio was 80:20, 10% of the training data was taken out for validation. We transformed the train, validation, and test sets into pairs of images before feeding the data to the models. The generated pairs are labeled either dissimilar (i.e., 0) or similar (i.e., 1). If both of the image samples belong to the same class then the label for the pair would be 1 and 0 otherwise. These pairs are then fed to the models. CNN-based Siamese neural network is our baseline model which yielded test accuracy of 79.67% with an AUC score of 0.87. Although VGG16 gave test accuracy upto 89.34%, but its fine tuning took more time than EfficientNetB0, thus, EfficientNetB0 was proven to be more time efficient and also lightweight than VGG16. The EfficientNetB0 substantially outperformed the other two models, with training and testing accuracy of 98% and 97%, respectively and AUC score of 0.97. Comparison on the basis of accuracy and AUC score is presented in Table 1. Furthermore, detailed comparison of the models using various performance measures such as recall, precision, and F1-score for similar and

Table 1 Performance comparison of efficientNetB0 with VGG16 and baseline model with respect to the accuracy and AUC score

Model	Training accuracy (%)	Testing accuracy (%)	AUC score
CNN-based Siamese model (baseline)	79.38	79.37	0.8750
VGG16-based Siamese model	92.29	89.34	0.9560
MetaEfficientNet (Proposed)	**98.23**	**97.23**	**0.9734**

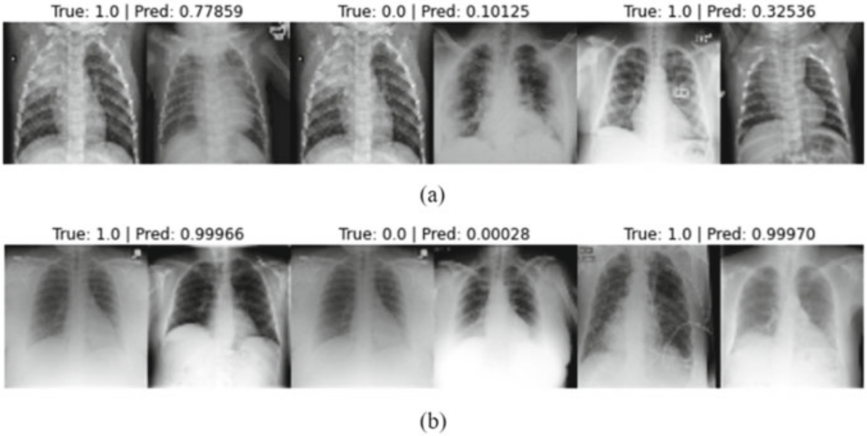
The [bold] in the tables are the proposed models that gave the best evaluation metrics score dissimilar pairs of images is given in Table 2. From Fig. 9, it can be observed that the proposed model can predict the similarity scores with greater precision when compared to the baseline model.

One of the criteria used to evaluate a model's performance is how effectively it can minimize the loss. We have used the Contrastive loss function for all the three Siamese neural nets. The CNN Siamese model could reduce the loss upto 0.1573 in training and 0.1545 in validation phase. The VGG16 performed slightly better as it was able to converge the loss upto 0.0646 in training and 0.0841 in the validation phase. However, the EfficientNetB0 model successfully minimized the training loss upto 0.0162 and validation loss upto 0.0380. According to the line graph shown in Fig. 10 the training loss of EfficientNetB0 appears to drop steadily throughout the training. Even though the validation loss exhibits a zig-zag line indicating little fluctuations from one epoch to another, the proposed model's validation loss appears

Table 2 Performance comparison of efficientNetB0 with VGG16 and baseline model based on recall, precision, F1-score for similar and dissimilar pairs of images

Model pair label	Recall		Precision		F1-score	
	Dissimilar	Similar	Dissimilar	Similar	Dissimilar	Similar
CNN based Siamese model (baseline)	0.68	0.91	0.88	0.74	0.77	0.82
VGG16 based Siamese model	0.92	0.92	0.92	0.92	0.92	0.92
MetaEfficientNet (Proposed)	**0.96**	**0.98**	**0.96**	**0.98**	**0.97**	**0.97**

The [bold] in the tables are the proposed models that gave the best evaluation metrics score


True: 1.0 | Pred: 0.77859
True: 0.0 | Pred: 0.10125
True: 1.0 | Pred: 0.32536
True: 1.0 | Pred: 0.99966
True: 0.0 | Pred: 0.00028
True: 1.0 | Pred: 0.99970

(a)

(b)

Fig. 9 Similarity score prediction, **a** Results of CNN-based Siamese net (baseline). **b** Results of EfficienNetB0 based Siamese net (Proposed)

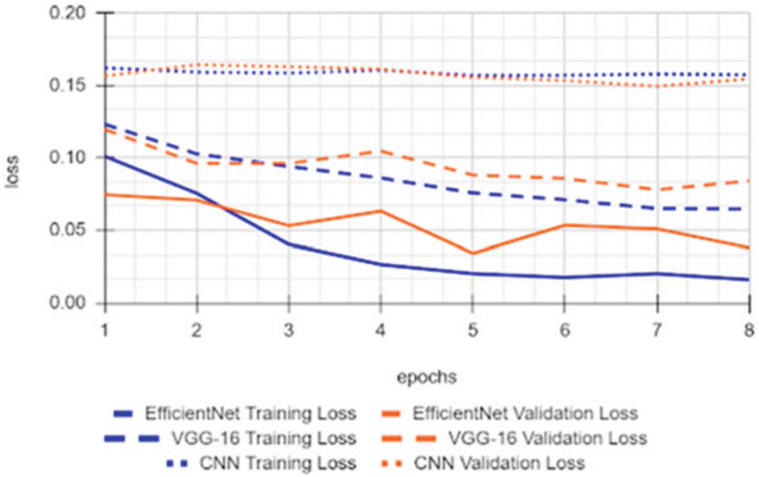


Fig. 10 Training and validation losses of CNN Siamese (baseline), VGG16 Siamese and MetaEfficientNet (Proposed)

to be minimum among the three. In contrast to EfficientNetB0, the line graph of CNN baseline is stable with almost negligible gradient.

5.3 3-Way-8-Shot Learning Testing with MetaEfficientNet

After fine tuning the proposed model on training data for classifying similar and dissimilar images, we took the samples from training data to build our support set. Based on our experimentation study, for testing we chose a 3-way-8-shot approach to classify the test data. Support set is the set of images which is used to compare against the incoming test image or query image (Fig. 11). For this study, we have used 8 images for each of the classes of “Normal”, “COVID-19”, and “Pneumonia” as the support set. Query set is the test set or test data which needs to be classified. We have used the support set to gain the average similarity of the test images with each of the support images from every single class. The average similarity value of the test image with the image in the support set of each class is calculated, after which the test image is classified as of that class which shows the highest similarity with that image.

While dealing with medical problems, the recall which is also known as sensitivity is one the most crucial performance measures for the model. It is used for those Machine learning and deep problems where the false negative rate has to be minimum. Recall is defined as the number of positive labels predicted out of all the actual number of positive labels, therefore it is also referred to as the true positive rate. Our 3-way-8-shot model yielded an average recall of 0.97. It gave the highest recall of 0.99 for

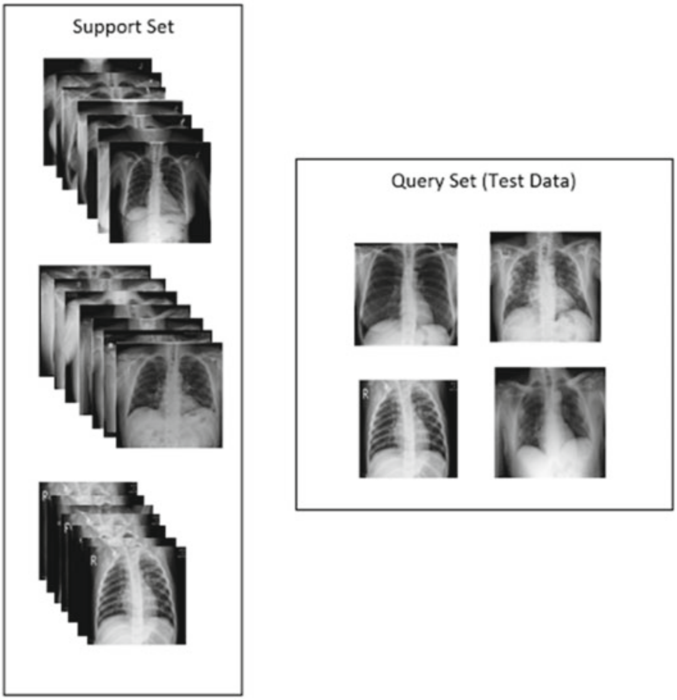


Fig. 11 Support set and query. Above figure illustrates the support set which consists of 8 samples of three classes for the 3-way-8-shot learning and a query set or the test set having randomly chosen images which are to be classified

the class Viral Pneumonia, followed by COVID-19 0.98, followed by Normal 0.96 and a testing accuracy of 98%. Table 3 shows the performance comparison of our robust MetaEfficientNet model for 3-way-k-shot with other state of the art methods

Table 3 Comparison of proposed model with state of the art methods

Reference no	Approach	Pre-trained model size	Accuracy	Precision	Recall	F1-score
[25]	Siamese + DenseNet121	33 MB	96.4%	0.965	0.962	0.959
[26]	Siamese + VGG16 (Meta COVID-19 3-way-10-Shot)	528 MB	96.5%	0.980	0.970	0.974
MetaEfficientNet	**Siamese + EfficientNetB0 (3-way-8-Shot)**	**29 MB**	**97.23%**	**0.980**	**0.976**	**0.976**

The [bold] in the tables are the proposed models that gave the best evaluation metrics score using metrics like precision, recall, and accuracy and also with respect to the sizes of the pre-trained models used.

6 Conclusion

Lung diseases are still among the most fatal diseases in the world. In 2020, we saw how COVID-19 was able to bring an entire planet to a halt. We may never be able to predict such circumstances, however, we can certainly prepare for such a drastic state of affairs. Diagnosis of a disease is one of the areas where we can improve tremendously. Current diagnostic tools, albeit accurate, are not very efficient and cheap. Advances in technologies have paved the way for faster diagnostic tools one of them being the use of machine learning and deep learning methods. Conventional methods of machine learning and deep learning have set impressive benchmarks but at the cost of the requirement of immense data. In this paper, we have successfully proposed a MetaEfficientNet model for the chest x-ray image classification. We performed a comparative analysis between our proposed model and baseline model CNN and VGG16, our proposed model successfully achieved an accuracy of 97%. Moreover, the experimentation results revealed that the MetaEfficientNet for 3-way-8-shot learning worked better than the state-of-the-art models.

References

1. Forum of International Respiratory Societies (2017) The global impact of respiratory disease—Second Edition. Sheffield, European Respiratory Society
2. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, Goo JM, Aum J, Yim JJ, Park CM (2019) Development and validation of a deep learning—based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 69:739–747
3. Tobias RR, De Jesus LCM, Mital MEG, Lauguico SC, Guillermo MA, Sybingco E, Bandala AA, Dadios EP (2020) CNN-based deep learning model for chest X-ray health classification using tensorFlow. In: Proceedings of the 2020 RIVF international conference on computing and communication technologies, RIVF 2020, Ho Chi Minh, Vietnam, 14–15 October 2020
4. Ahsan MM, Alam TE, Trafalis T, Huebner P (2020) Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients. Symmetry 12
5. Kieu STH, Bade A, Hijazi MHA, Kolivand H (2020) A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions. J Imaging 6(12):131. https://doi.org/10.3390/jimaging6120131
6. Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW (2020) Efficient Pneumonia detection in chest X-RAY images using deep transfer learning. diagnostics (Basel). 10(6):417. Published 2020 Jun 19. https://doi.org/10.3390/diagnostics10060417
7. Militante SV, Dionisio NV, Sibbaluca BG (2020) Pneumonia detection through adaptive deep learning models of convolutional neural networks. In: 2020 11th IEEE control and system graduate research colloquium (ICSGRC), Shah Alam, Malaysia, pp 88-93. https://doi.org/10.1109/ICSGRC49013.2020.9232613

8. Mostofa A, Rahul G, Anne D (2019). Application of a convolutional neural Network using transfer learning for tuberculosis detection. pp 427–433. https://doi.org/10.1109/EIT.2019.8833768
9. Ravichandiran S (2018) Hands-on meta learning with python. In: Ramchandani P et al. (ed) Packt Publishing Ltd, December 2018. www.packtpub.com
10. Jadon S (2020) An overview of deep learning architectures in few-shot learning domain. ArXiv, abs/2008.06365
11. Li MD, Chang K, Bearce B et al. (2020) Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. npj Digit. Med 3:48. https://doi.org/10.1038/s41746-020-0255-1
12. Shorfuzzaman M, Shamim Hossain M (2021) MetaCOVID: a Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. Pattern Recogn 113:107700. ISSN 0031-3203. https://doi.org/10.1016/j.patcog.2020.107700
13. Prayogo K, Suryadibraya A, Young J (2020) Classification of pneumonia from x-ray images using siamese convolutional network. Telkomnika (Telecommunication Computing Electronics and Control) 18(3):1302–1309
14. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Al-Emadi N, Reaz MBI, Islam MT (2020) Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8:132665–132676
15. Daniel K, Kang Z, Michael G (2018) Labeled optical coherence tomography (OCT) and chest X-ray images for classification. Mendeley Data V2. https://doi.org/10.17632/rscbjbr9sj.2
16. Ravichandiran S (2018) In: Hands-on meta learning with python meta learning using one-shot learning, MAML, reptile, and meta-SGD with TensorFlow
17. Karen S, Andrew Z (2014) Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556
18. Sitaula C, Hossain MB (2021) Attention-based VGG-16 model for COVID-19 chest X-ray image classification. Appl Intell 51:2850–2863. https://doi.org/10.1007/s10489-020-02055-x
19. Kandel I, Castelli M (2020) The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT Express 6(4):312–315. ISSN 2405–9595. https://doi.org/10.1016/j.icte.2020.04.010
20. https://neurohive.io/en/popular-networks/vgg16/
21. Mingxing T, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, PMLR, pp 6105–6114
22. Marques G, Agarwal D, de la Torre Díez I (2020) Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. Appl Soft Comput 96:106691. ISSN 1568-4946https://doi.org/10.1016/j.asoc.2020.106691
23. Duong LT, Nguyen PT, Di Sipio C, Di Ruscio D (2020) Automated fruit recognition using EfficientNet and MixNet. Comput Electron Agricul 171:105326. ISSN 0168-1699. https://doi.org/10.1016/j.compag.2020.105326
24. Tryan P, Syahidah R, Jenq-Shiou L (2020). Enhanced skin condition prediction through machine learning using dynamic training and testing augmentation. IEEE Access 1–1. https://doi.org/10.1109/ACCESS.2020.2976045
25. Shruti J (2021). COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach. 1. https://doi.org/10.1117/12.2581496
26. Mohammad S, Shamim Hossain M (2021) MetaCOVID: a Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. Pattern Recogn 113:107700. https://doi.org/10.1016/j.patcog.2020.107700

Chapter 20
Application of Deep Machine Learning Techniques in Oil Production Forecasting

Tushar S. Lanjekar and Hrishikesh K. Chavan

1 Introduction

Production oil forecasting can impact economics of the oilfields and planning to generate the future scope. Conventional tools such as decline curve are used by reservoir and production engineer manually for such analysis and prediction to get the exact oil rate estimation [1]. Traditional techniques such as decline curve analysis are not flexible in handling complexities and hidden parameters of the reservoir [2]. They might take long hours to process if oil production data are large. Above mentioned limitations can be addressed by machine learning methods. Therefore, it has been provided opportunity through growing and techniques which used are artificial intelligence (AI) and machine learning (ML) [3, 4].

Data-driven approach of the usage of network by recurrent neural (RNN) and neural network by convolution (CNN) has proven to be highly effective and accurate in applications related to time series prediction of stock market and oil price previously [5–7]. These ML techniques have been used to forecast production of 5 oil wells.

The methods and basic logical sequences from (Brownlee 2018) predicting using LSTMs were implemented [8–10]. The field of study and the proposed networking is studied based on the previous data, and then, a structural model is developed to predict the required time data. Network that is suitable for processing oil and future prediction is long short-term memory (LSTM) and prediction dataset with errors and delays in a time collection. Based on complicated and accountable historic facts set, an accountable time collection prediction technique primarily based totally on a long-time period reminiscence cycle and a neural community is used [11–13].

T. S. Lanjekar (✉) · H. K. Chavan
School of Petroleum Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_20

The CNN-based and RNN and long short-term memory (LSTM) model is created using various input and output sequences. DCA is taken as result comparing conventional tool. For number of wells, appropriate input parameters were pooled together before projections were rejected using the previous model. With various commands and continuous training, the validation is performed, observed with the aid of using an assessment of version accuracy and efficiency. Related research represents that CNN is appropriate for processing time series data samples. Besides the types of oil production data, this paper proposes an oil forecasting structure on CNN-LSTM that can effectively predict oil production with upcoming time portion and offers a robust foundation for formulating reasonable oil production results [9, 14, 15].

2 Methodology

Steps provided to execution of the result by means of data collection, filtration, and algorithm application.

2.1 Conventional Tool Analysis by Decline Curve

The decline curve involves change in oil flow rate over the period of time. This is an oil production forecast technique which used to predict behavior of oil production with respect to time. Initially, decline curve uses production with historical data available volumes then extrapolates the curve to forecast oil production. Equation 1 is a general equation for decline curve analysis given by (Arps 1944) [1].

$$q(t) = q_i/(1 + bd_i t)^{\frac{1}{b}} \tag{1}$$

where

q = Production rate current, q_i = Production rate initial,

d_i = d = dt = Decline rate nominal, t Production time for cumulative oil production, b = (0–1) Decline constant hyperbolic.

The decline curves are divided in to three categories based on decline constant b:

- Exponential

  In case of exponential decline, the value of b, i.e., decline constant is equal to 0. After substituting the value of b in Eq. 1, we get

$$q = q_i \mathrm{e}^{-\mathrm{d}t} \tag{2}$$

- Hyperbolic

  In case of hyperbolic decline, the value of b is in the range of 0 to 1. Hence, the Eq. 1 becomes as follows.

$$q = q_i/(1 + bd_i t)^{\frac{1}{b}} \tag{3}$$

- Harmonic

The value of b is 1 in case of harmonic decline. After using this value of b, Eq. 1 is reduced to

$$q = q_i/(1 + d_i t) \tag{4}$$

2.2 RNN—Recurrent Neural Network

The oil prediction made using DCA is a smooth curve that cannot capture the rapid variations that usually happen in the production process and is therefore subject to a certain amount of error, nevertheless, engineers still favor this method due to its simplicity. Oil production flow rate is kept on changing creates reading and forecasting error. Deep learning architectures along with clustering to adaptively integrate the various sources of information and features relevant to the forecast. This neural network is based on values provided average, and output layer is involved. Interconnected nodes are involved in this network to each layer. Nodes between the layers are disconnected [6, 8, 16]. As shown in the formula, X indicates the value of input layer which is related to oil production flow at that particular moment, S represents the value of hidden layer which will relate the oil production value with other nearby values, weight matrix is used "U" for hidden layer which has been came from input layer, whereas the output is predicted flow for the next time period is "O." Similarly, to minimize error the weight matrix is used "V" which is for next output from input. Last input parameter is "W" of hidden layer. The error minimization factor at "t" and "t–1" is applied by (Eq. 5)

$$S_t = F(Ux_t + Ws_{t-1}) \tag{5}$$

Recurrent neural network (RNN) wherein connections among various series are considered temporary. Feedforward network has been established to internal nation to variables of track and steps. Together should have more stored and compacted states, direct control thru manner of manner of the neural. Such managed states are known as a gated nation or gated reminiscence and are a part of lengthy short-time period reminiscence networks (LSTMs) and gated recurrent devices. The area and type of networks through by RNN network within and among nodes shape represented by the pictorial graph with a temporal series [17]. This permits it to showcase temporal dynamic behavior. The actual performance segregated from next go ahead step neural networks, RNNs can use their inner nation (reminiscence) to system variable duration sequences of inputs. This ex- amination proposed a singular

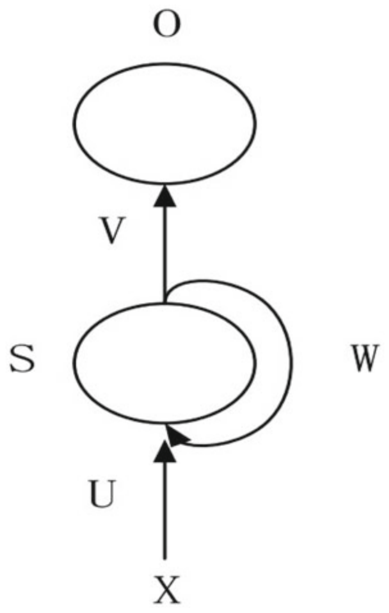


Fig. 1 Structural diagram of recurrent neural network [17]

method to version time collection associated problems (e.g., production forecast) the usage of the RNN primarily-based totally series-to-series models (Fig. 1).

2.3 CNN—Convolutional Neural Network

CNN, even though popular in picture datasets, also can be used (and perhaps more realistic than RNNs) on time series statistics. The spatial volume of this connectivity is known as the receptive area of the node. The neighborhood connectivity is finished via replacing the weighted sums from the neural network with convolutions. In each layer of the convolutional neural network, the center is convolved with the load matrix (additionally known as the clear out) to create a feature map. In other phrases, the load matrix slides over the enter and computes the dot product among the center and the burden matrix. Note that as against ordinary neural networks, all of the values inside the output characteristic map proportion the identical weights. This manner that all the nodes within the output hit upon exactly the identical sample [18, 19] (Fig. 2).

The input sample duration is defined as l, the convolution kernel length is *c*, and the down sampling amplitude is *S*.

$$[(l - c + 1)/s - c + 1/\mathrm{s} = \mathrm{N}] \tag{6}$$

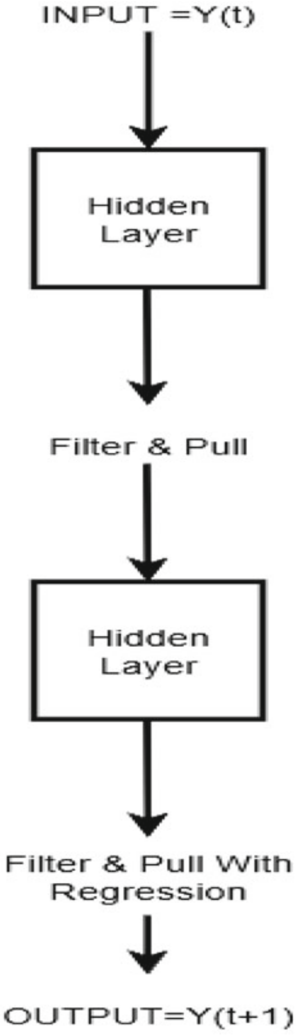


Fig. 2 Basic architecture of convolution neural network [6]

While the convolution layer and the down sampling layer are constant to two, the above 3 parameters have to satisfy the following members of the family. When the community structure consists of most effective one-layer convolution and down sampling, the above three parameters need to satisfy the subsequent members of the family in the above formulas, N is a effective integer, no longer equal to at least one. The size of the convolution kernel and the amplitude of the down sampling are limited by using the period of the enter samples. For the economic time series records, how lengthy the fee of a positive time impact relies upon at the traits of the monetary time collection records after which affects the length of the input pattern. The variety of convolution core size similar to the unique input pattern lengths while the community structure contains layers of convolution, down sampling, and down sampling amplitude of 2.

2.4 LSTM—Long Short-Term Memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. The non-relevant values of oil production impacts on irrelevant forecasting such as on few days oil production data are missing or the drastic change in oil production quantity for a particular day may impact on data prediction.

For this reason, the LSTM is an improvised structure of step by step data analysis with the recurrent neural network and convolution neural network cannot handle, i.e., long-distance dependencies [6, 9]. Figure 3 is a schematic structural view of the LSTM, where $f(t)$, $i(t)$, $C(t)$, and $O(t)$ are the forget gate, the input gate, the output gate, and the timing of the output gate, respectively, and W, b, and tanh are the corresponding weights which provide deviations and excitation functions which reduce and start the new command with error minimization for next oil production flow rate, respectively [8, 20, 21].

The forget gate shows the value of $C(t–1)$ at the value of the previous stage, and it is placed at present stage $C(t)$; $X(t)$ shows the value present at the input stage which

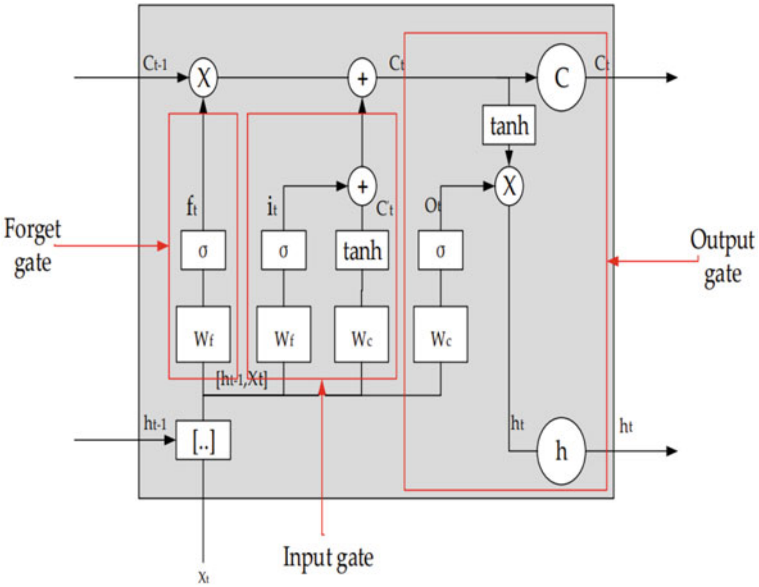


Fig. 3 Structural diagram of LSTM [17]

is related to current value and of the network is saved to the unit state $C(t)$ at the current moment and the state which controls the value $C(t)$ is output to the LSTM. The current output value is ht. Its calculation formula is shown by Eqs. (7–9). The forget gate *ft* is Eq. (7) the input gate, it is Eq. (8), and the output gate $O(t)$ is Eq. (9)

Long short-time period absorption by the previous data in LSTM is a pretended recurrent neural cell block (RNN) structure is used with inside the discipline of deep studying. Unlike popular feedforward neural networks, LSTM has comments connections. It can system now no longer most effective unmarried statistics points (consisting of images), however, additionally whole sequences of statistics. A not unusual place LSTM is molecular-based unit, an execution gate, a result gate, and an overlook gate. Algorithm recollects parameters on arbitrary sequential time durations and the 3 gates alter the float of records into and out of the molecular (Fig. 4).

$$f_t = \delta(\omega_f.\left[h_{t-1}, x_t\right] + b_f \tag{7}$$

$$\begin{aligned} i_t &= \delta(\omega_i.\left[h_{t-1}.x_t\right] + b_i \\ C^1 &= \tanh(\omega_c.\left[h_{t-1}, x_t\right] + b_c \\ c_t &= f_t.c_{t-1} + i_t c_t' \end{aligned} \tag{8}$$

$$\begin{aligned} o_t &= \delta(\omega_0 \cdot \left[h_{t-1}, x_t\right] + b_0 \\ h_t &= O_t \cdot \tanh(c_t) \end{aligned} \tag{9}$$

Oil production flow rate varies with time period for day to day note. It creates delay to understand the flow rate for triggering function. This delay is converted to error for prediction and accuracy of next day oil flow prediction and forecast. To

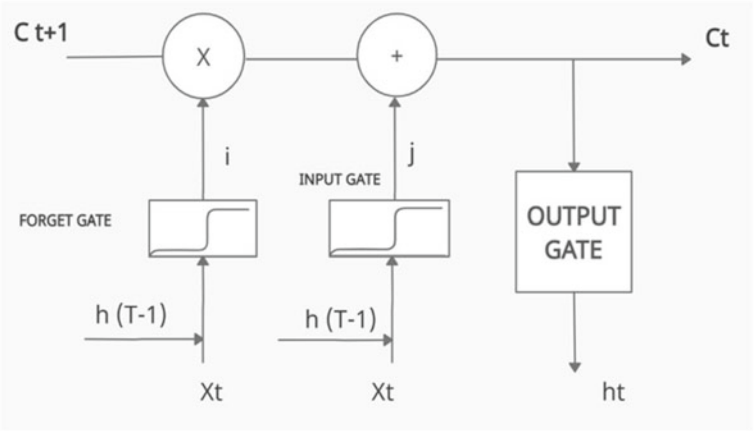


Fig. 4 Structural diagram neural network [14]

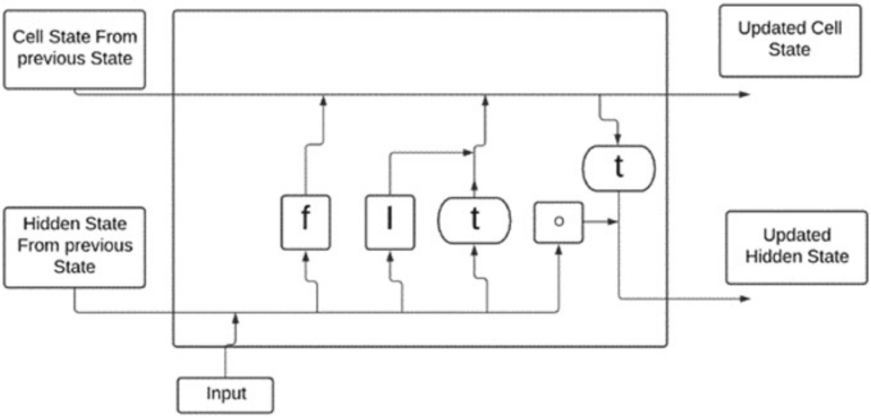

Cell State From previous State
Updated Cell State
t
f
I
t
o
Hidden State From previous State
Updated Hidden State
Input

Fig. 5 Design and flow of long short-term memory [22]

overcome this scenario, the LSTM is used by supplying and minimizing the error of the previous oil prediction data.

Input Gate: Oil production previous data are the input valve and parameter for the sequential step. The functions with time and the memorized value from state previous to the current are considered for the forward logical value of oil production rate. Sometimes the value may be 0 or non, which create delay and error to understand the pattern of the previous oil production rate which creates error to further prediction of oil forecasting. This error is minimized by LSTM steps (Fig. 5).

Execution gate: The output execution gate makes use of an enter much like the preceding gates. An aggregate of tanh activation implemented to the molecular state long-time period reminiscence and values obtained molecular output is the reminiscence-molecular, the horizontal line going for walks thru the pinnacle. It is just step by step intervention. Actual treating quantity of statistics important to compute the version output. It proceeds directly down the complete sequence, with just a few minor linear interactions. It is very smooth for statistics to simply go with the drift alongside it unchanged. This idea realizes long short-time period memory. LSTM having very initial stage to start is what statistics we are going to throw far from the molecular state. The input gate layer takes crucial records into the reminiscence-molecular to keep during learning. Such out-put will be based mostly on our molecular nation, but will be a filtered version. First, we run a sigmoid layer which makes a choice what additives of the molecular nation we are going to output. Then, we positioned the molecular nation via tanh (to push the values to be between 1 and 1) and multiply it through manner of way of "the output of the sigmoid gate," simply so we high-quality output the additives we decided to. A traditional RNN primarily-based totally lengthy short-time period memory (LSTM) version becomes first advanced with diverse enter and output sequences. Then, famous DCA had been applied as reference solutions. Moreover, the facts cleansing method entails the coaching of records production fees and nicely constraints for current wells. Finally,

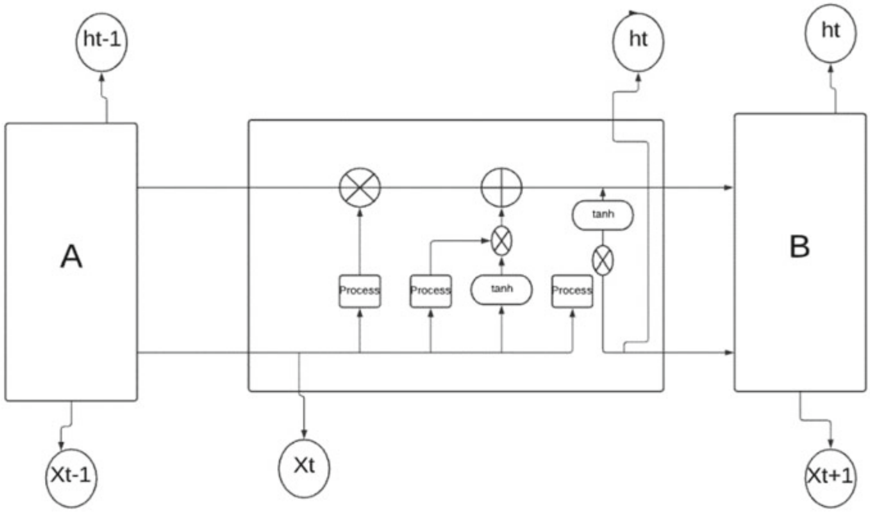


Fig. 6 LSTMs the memory-cell [24]

hold-out schooling and validation had been done, observed via way of means of an assessment of version accuracy and efficiency. Various LSTM primarily-based totally series-to-series fashions along with one-to-one, many-to-one, and many-to-many had been correctly applied for production forecast [23] (Fig. 6).

3 Well Production Description

The oil production data from two different Indian fields collected to assess application of CNN, RNN, and LSTM for oil forecasting. Well-1, well-2, and well-3 are from Camay basin field X which falls under Category-I, i.e., proven reserves. Well-4 and well-5 are from Category-II basin field Y which includes basin with contingent reserves. The description of all oil wells is as follows:

Well-1

The data have been collected from five well data from an Indian oil fields to analyzes the profile and forecast. Initially, well-1 data are studied significantly to get the required results resented in the (Fig. 7). As represented diagram shows the initially oil production rate was very high and the natural pressure profile, it went down day by day and plotted with respect to the outcome of oil production quantity. Data are very noisy for the period of year 2016 to 2017. There is continuous decline in production rate since 2018.

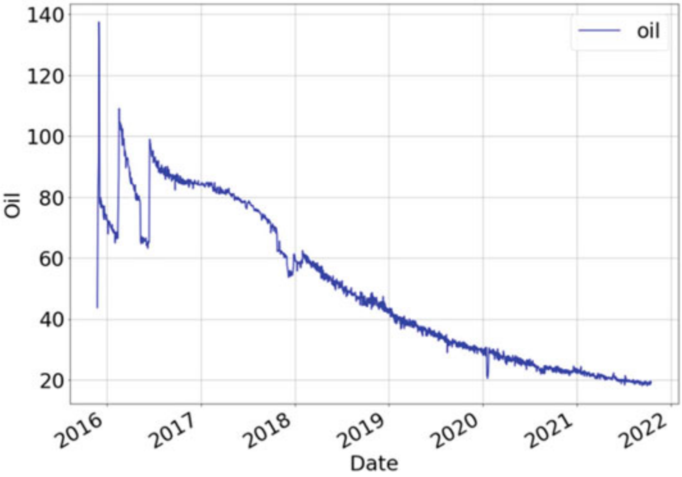


Fig. 7 Well-1 production profile

Well-2

Figure 8 shows oil production profile for well-2. At the start of production, well-2 achieved more than 2100 bbl/day oil flow rate. By the start of 2017, oil rate decreased to almost 100 bbl/day. Oil production rate consolidated and steadied around 800 bbl/day in late 2018, and it is declining since 2019.

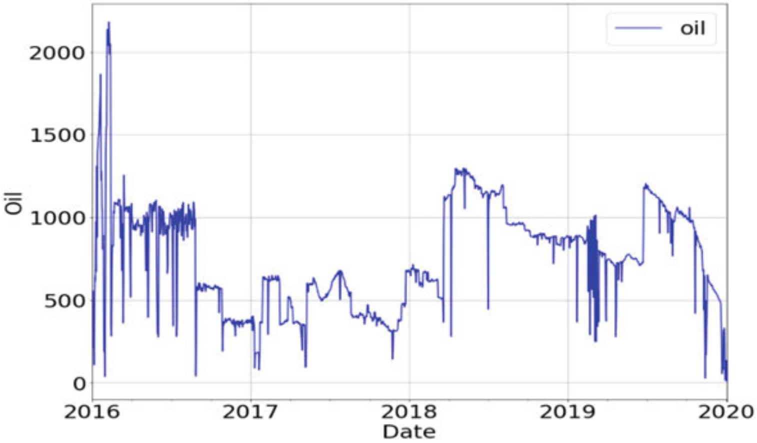


Fig. 8 Well-2 oil production profile

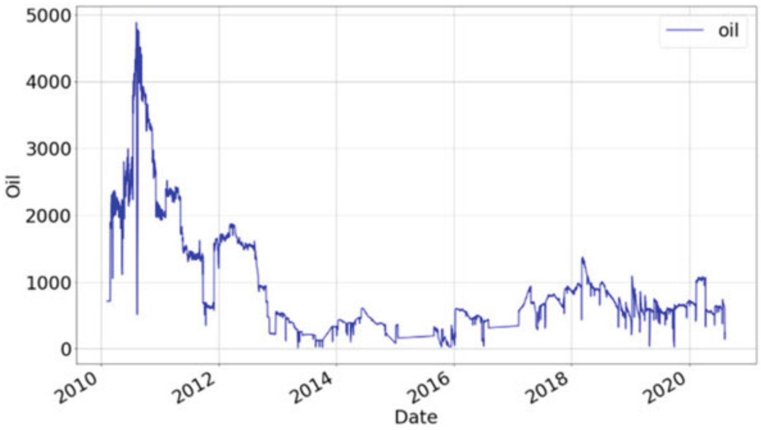


Fig. 9 Well-3 production profile

Well-3

It has been observed that from (Fig. 9) that data are value with zero on particular dates. Due to this, the graph is not relevant. Such cases in machine learning, the values are removed or the average has been taken to normalize the graph year 2013 to 2017 the values are with zero rate of oil production. It has been a unique graph with actual variations in oil production. Graphical representation values from year 2018 to 2019 are closely represented.

Well-4

This data are from the year 2011 to 2020. The production of the well is very high rate as it represented in (Fig. 10). It shows that the changes are occurred in production

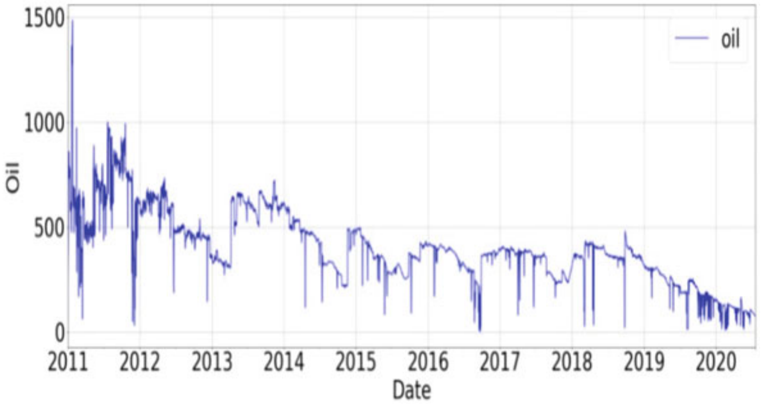


Fig. 10 Well-4 production profile

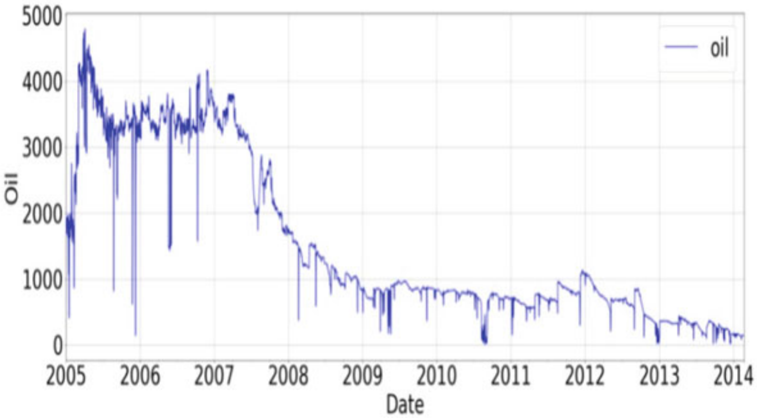


Fig. 11 Well-5 production profile

from the year 2013 to 2019 was consistent, but from 2019 year, the oil production has been drastically changed.

Well-5

The result from another four dataset is appropriate to the well-1 observations to get the actual profile for the dataset. Dataset of well-3 is cleary noted about well-5 is with error as the values of particular dates are not available. Here after the implementation of machine learning is inserted on the graphical representation of forecasting (Fig. 11).

4 Result and Discussion

4.1 Production Forecasting using RNN

RNN is applied and model has been created for well-1 data to understand the train and test with valid forecasting. 20% of the data previous one is used to set the ideal model for forecasting, and the results are appropriate to the developed structure. The evolved data-pushed method makes the system of records matching and forecasting performance and correct for belongings without or with first rate operation records information (Figs. 12, 13, 14, 15 and 16).

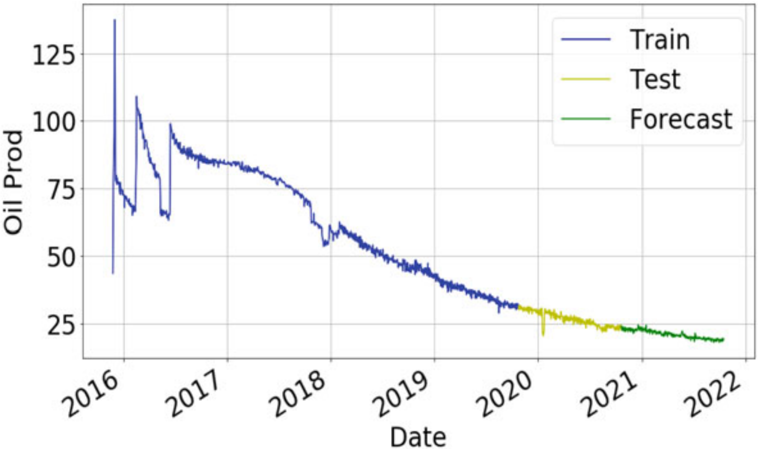


Fig. 12 Application of RNN to forecast well-1 oil production

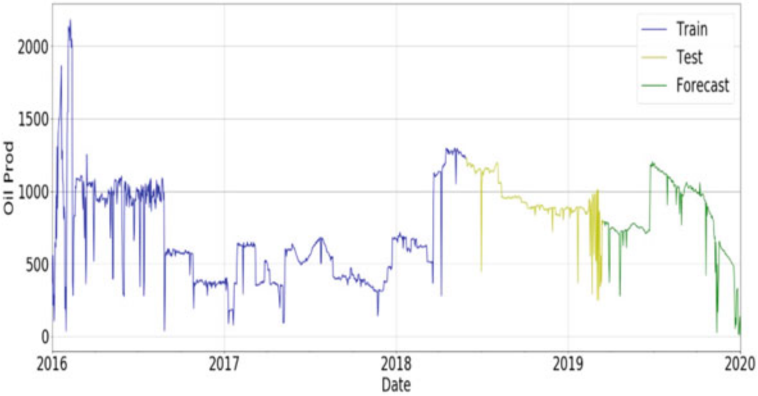


Fig. 13 Application of RNN to forecast well-2 oil production

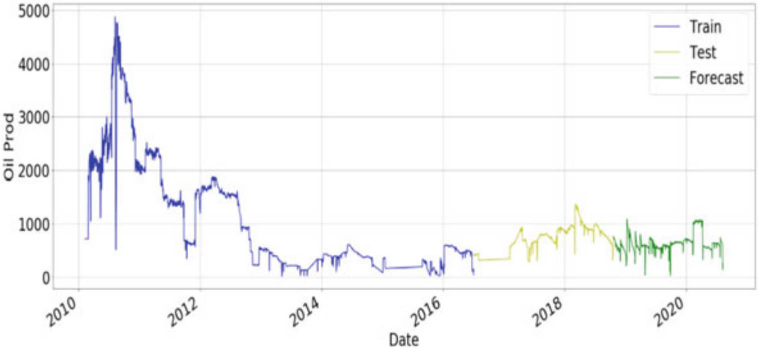


Fig. 14 Application of RNN to forecast well-3 oil production

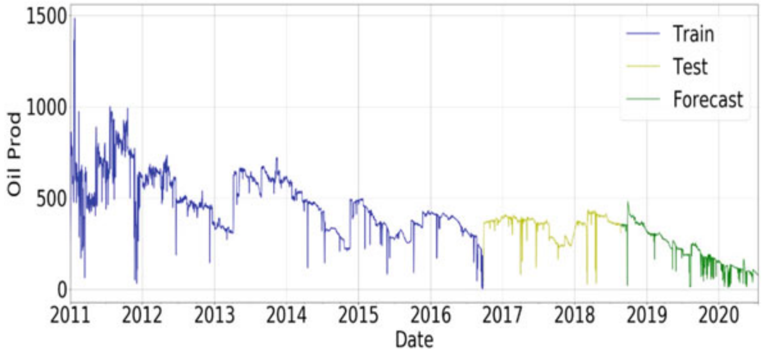


Fig. 15 Application of RNN to forecast well-4 oil production

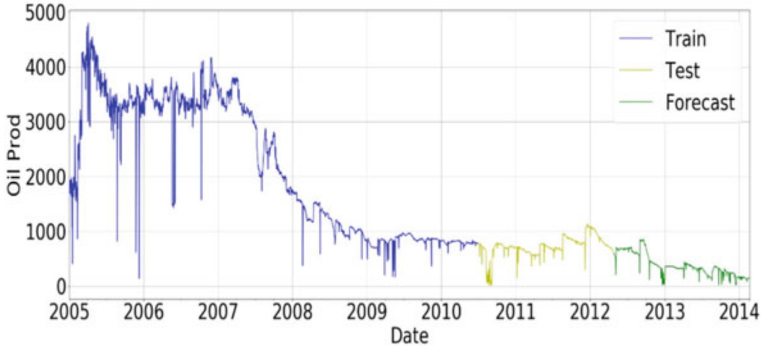


Fig. 16 Application of RNN to forecast well-5 oil production

4.2 Production Forecasting by RNN and LSTM

RNN + LSTM neural network are applied and model has been created for well-1 data to understand the train and test with valid forecasting. 20% of the data previous one is used to set the ideal model for forecasting, and the results are appropriate to the developed structure. 60% data are considered as train set, and 20% data are for test set. The evolved data-pushed method makes the system of records matching and forecasting performance and correct for belongings without or with first rate operation records information. Problem faced by “gates” of RNN are replaced and corrected by use of LSTM model (Figs. 17, 18, 19, 20 and 21).

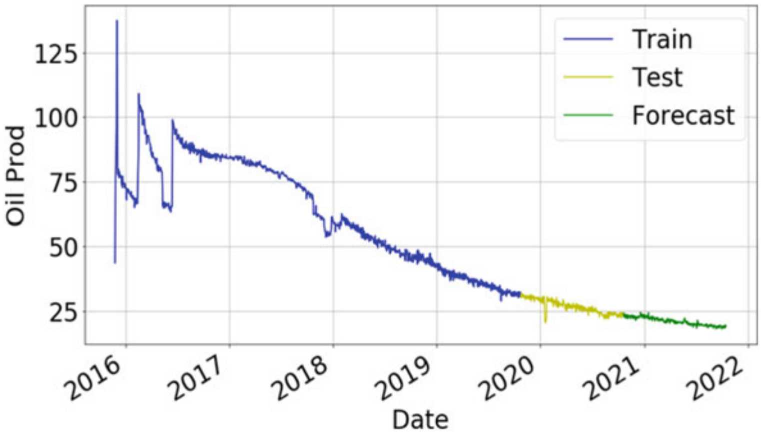


Fig. 17 Application of RNN + LSTM to forecast well-1 oil production

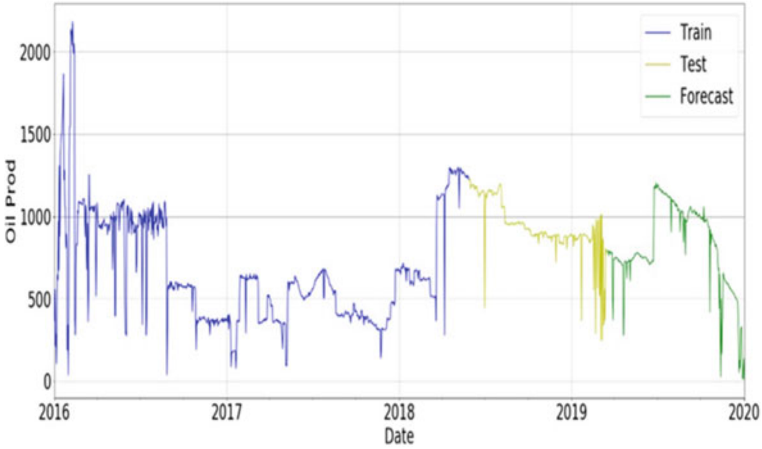


Fig. 18 Application of RNN + LSTM to forecast well-2 oil production

4.3 Production Forecasting using CNN

CNN is applied and model has been created for well-1 data to understand the train and test with valid forecasting. 20% of the data previous one is used to set the ideal model for forecasting, and the results are appropriate to the developed structure. It may be correctly implemented to the future time dependable series. The nearby belief sharing of CNN can substantially lessen the wide variety of parameters, hence enhancing the performance of version learning. The CNN version will study a feature that represents a series of beyond observations as enter to an output observation. First, it has been applied to well-1. Then, it has been applied to well-2,3,4,5 datasets. Dataset from

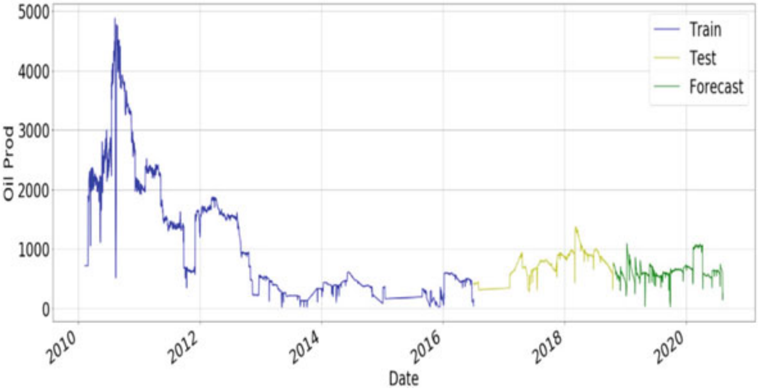


Fig. 19 Application of RNN + LSTM to forecast well-3 oil production

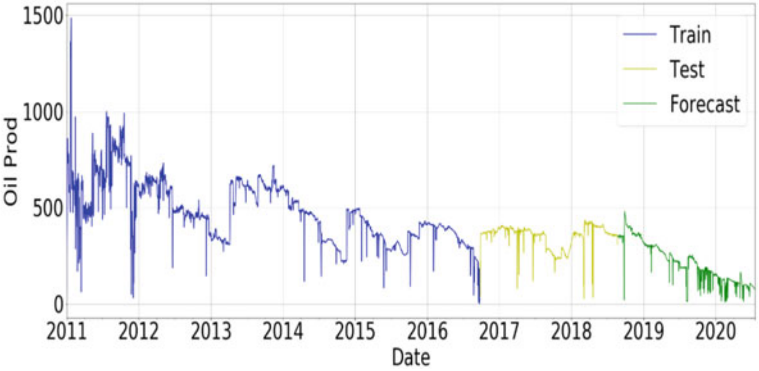


Fig. 20 Application of RNN + LSTM to forecast well-4 oil production

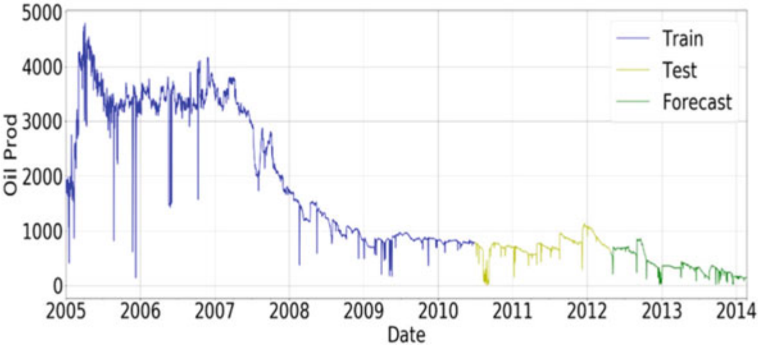


Fig. 21 Application of RNN + LSTM to forecast well-4 oil production

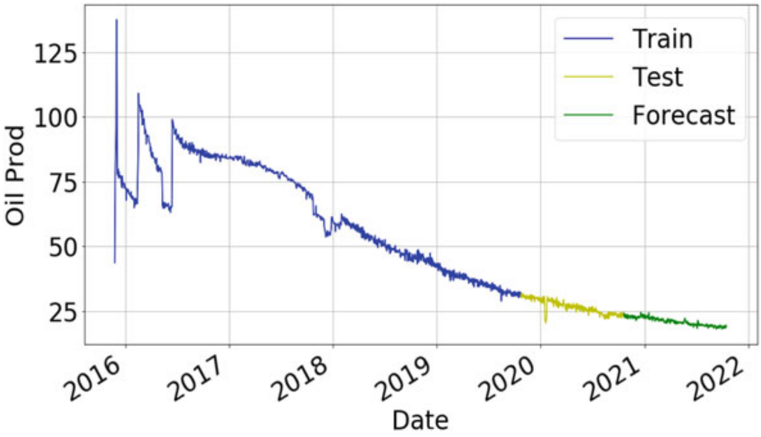


Fig. 22 Application of CNN to forecast well-1 oil production

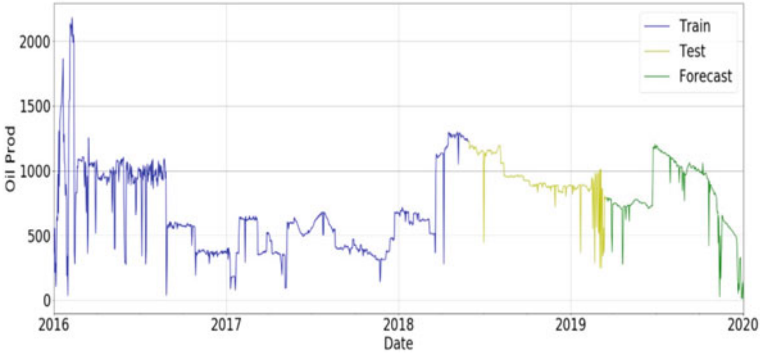


Fig. 23 Application of CNN to forecast well-2 oil production

(Fig. 1) is observed to be accurate for the time series forecasting of oil production. Further, it is applied to CNN. The train and test part are validated with good accuracy for convolutional neural network (Figs. 22, 23, 24, 25 and 26).

4.4 MSE Error Calculation

Mean square error is calculated based on the following equation as the whole concept of assessment index by perfection and accuracy:

$$MSE = \frac{1}{n}\Sigma_{\mathrm{i}-1}^{n}((f_i - y_i)^2 \tag{10}$$

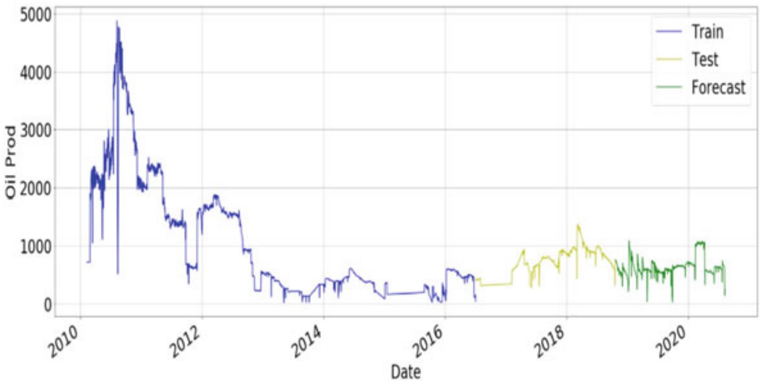


Fig. 24 Application of CNN to forecast well-3 oil production

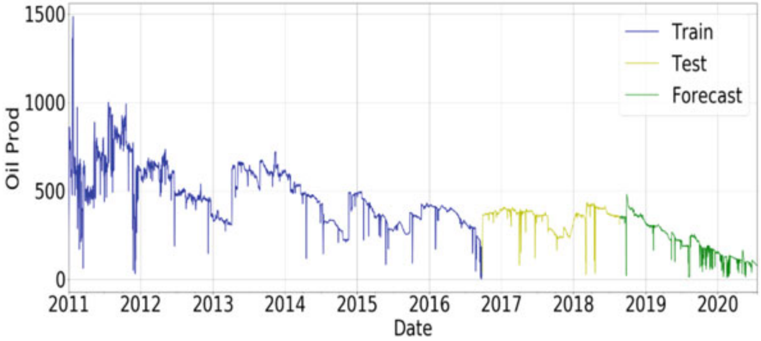


Fig. 25 Application of CNN to forecast well-4 oil production

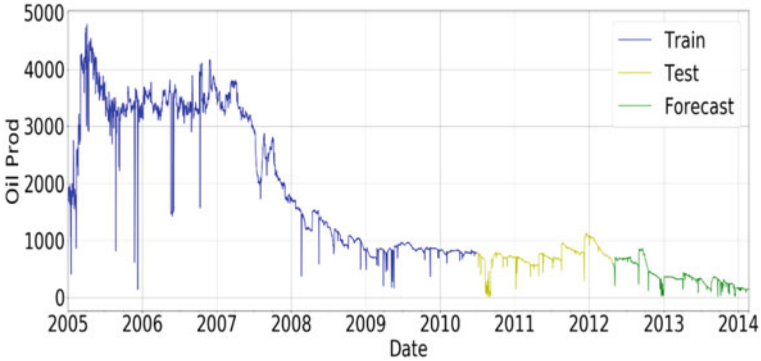


Fig. 26 Application of CNN to forecast well-5 oil production

Table 1 Validation RMS error for CNN, RNN, RNN, and LSTM to forecast for five oil wells

	RNN	RNN and LSTM	CNN
Well-1	1.327	1.23	0.918
Well-2	1.0802	1.102	1.002
Well-3	4.207	4.12	2.21
Well-4	3.212	3.111	2.201
Well-5	10.67	10.69	8.161

in which f_i, y_i are the expected fee and genuine fee, respectively. The take a look at consequences and actual consequences of every optimization version have been in comparison with assess the right impact of the version. The going for walks time of every optimization version in the course of the schooling take a look at become counted to assess the computational performance of the version.

The forecasting accuracy is normally calculated by the method of root mean squared error (RMSE) [25, 26]. It is given by following formula.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{\mathrm{i}-1}^{n}((f_i - y_i)^2} \tag{11}$$

The RMSE error is calculated to check feasibility of machine learning methods. Table 1 shows RMS error calculation for CNN, RNN, RNN, and LSTM.

4.5 Comparison of Decline Curves with CNN

As per Table 1, it is evident that CNN is showing highest accuracy among all other deep ML forecasting techniques. The accuracy of CNN is again checked with conventional tool decline curve analysis. Figures 27, 28, 29, 30 and 31 show that CNN is showing better fitting curve than decline curve analysis. To confirm this, error analysis of CNN and DCA has also been done. The results are shown in Table 2.

According Table 2, the accuracy of CNN is much better than decline curve analysis. The RMSE error of validation data of well-1 is as low as 0.918 and it goes up to 8.161 for well-5 where data are very noisy.

4.6 Time Taken by Each Algorithm to Complete the Process

As shown in Table 3, total run time for the data has been calculated. CNN is performing better other ML techniques and decline curve analysis in case of total time taken to run the whole data. CNN is faster than RNN, RNN and CNN, and DCA to forecast the oil production data.

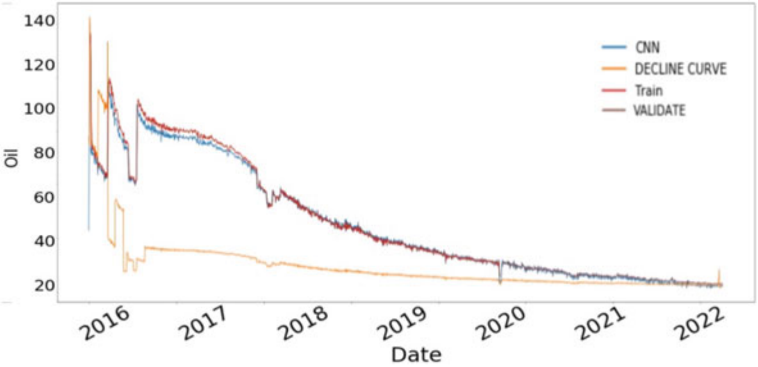


Fig. 27 Comparison of CNN and decline curve to understand better forecasting technique for well-1

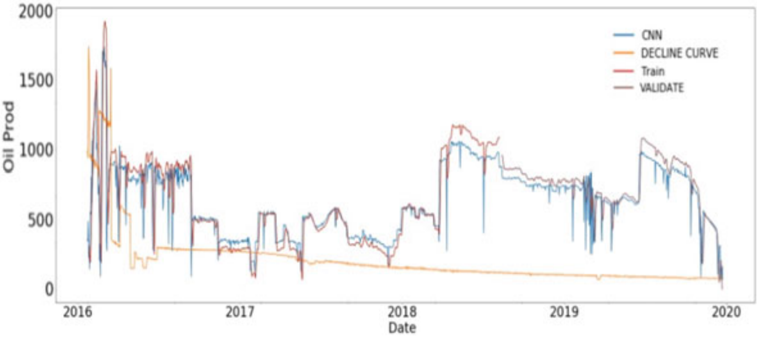


Fig. 28 Comparison of CNN and decline curve to understand better forecasting technique for well-2

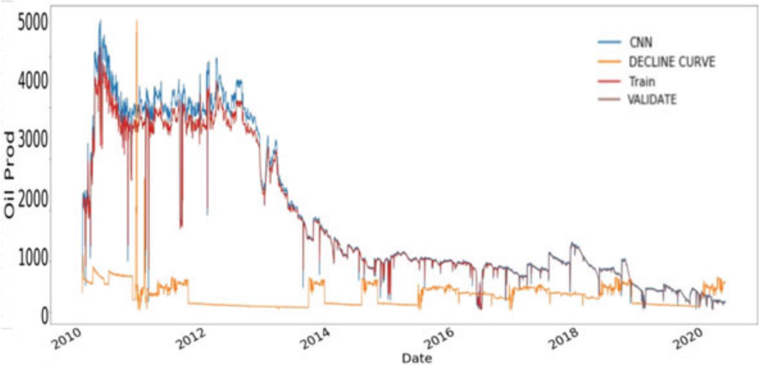


Fig. 29 Comparison of CNN and decline curve to understand better forecasting technique for well-3

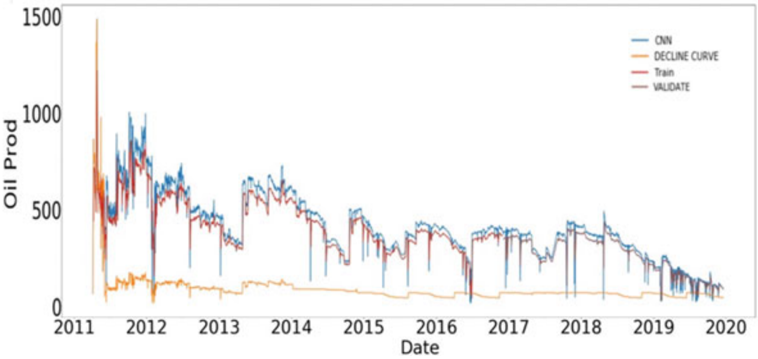


Fig. 30 Comparison of CNN and decline curve to understand better forecasting technique for well-4

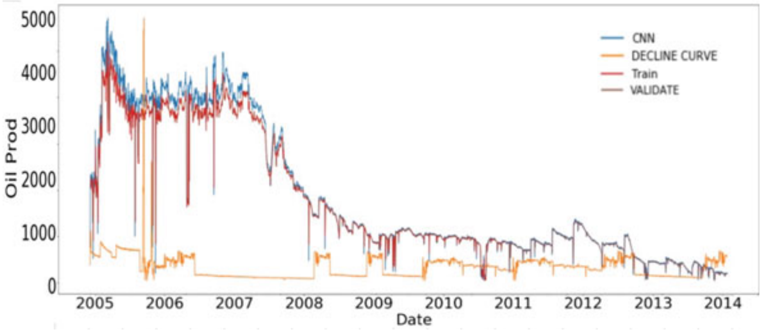


Fig. 31 Comparison of CNN and decline curve to understand better forecasting technique for well-5

Table 2 Comparison validation error of CNN and DCA to forecast for five oil wells

CNN	DCA
0.918	5.61
1.0002	5.326
2.21	6.712
2.201	4.121
8.161	11.112

5 Conclusion

Result tables clearly show that machine learning techniques can match and outperform conventional tools such as decline curve analysis while forecasting oil production. The consequences from those simulation research display that ML techniques have more suitable forecasting functionality with better accuracy within side the

Table 3 Comparison of validation error of CNN and DCA to forecast for five oil wells

		Total run time (Sec)			
	Total days	RNN	RNN and LSTM	CNN	DCA
Well-1	1451	34	32	24	60
Well-2	1171	115	105	80	58
Well-3	3614	121	114	95	129
Well-4	3791	180	179	122	162
Well-5	3672	210	152	102	112

prediction of oil production. However, the CNN offers oil production forecasting with the best prediction accuracy. CNN is also faster other ML techniques mentioned in this study. It shows that it can handle large amount of oil production data with better accuracy.

Acknowledgements Authors would like to thank Dr. Vishwanath Karad MIT World Peace University for allowing us to participate in the 3rd Springer International Conference on Computational Science and Applications 2021 (ICCSA2021) which is organized by School of Computer Engineering, MIT-WPU. We would also like to extend thanks to School of Petroleum Engineering for the support.

Author's Contribution
Tushar and Hrishikesh both are involved in this paper. Tushar designed ML models using Python. Hrishikesh analyzed the result. Tushar and Hrishikesh read and gave consent to this paper.

References

1. Arps JJ (1944) Analysis of decline curves. https://doi.org/10.2118/945228-G
2. Centilmen A, Ertekin T, Grader AS (1999) Applications of neural networks in multiwell field development. Proc-SPE Annu Tech Conf Exhib 1 (PI):33–43. https://doi.org/10.2523/56433-ms
3. Cao Q, Banerjee R, Gupta S et al (2016) Data driven production forecasting using machine learning. Soc Pet Eng—SPE Argentina Explor Prod Unconv Resour Symp. https://doi.org/10.2118/180984-ms
4. Ristanto T (2018) Machine learning applied to multiphase production problems. pp 1–71
5. Gupta S, Fuehrer F, Jeyachandra BC (2014) Production forecasting in unconventional resources using data mining and time series analysis. Soc Pet Eng—SPE Can Unconv Resour Conf 2014 1:247–254.https://doi.org/10.2118/171588-ms
6. Lu W, Li J, Li Y, et al (2020) A CNN-LSTM-based model to forecast stock prices. Complexity 2020https://doi.org/10.1155/2020/6622927
7. Moitra N, Raj P, Saxena S, Kumar R (2020) Crude oil prediction using Lstm. Int J Innov Sci Res Technol x:2016
8. Brownlee J (2020) Introduction to time series forecasting with python. Mach Learn Mastery 148:148–162
9. Brownlee J (2017) Long short-term memory networks with python. Mach Learn Mastery With Python 1:228

10. Brownlee J (2017) Master machine learning algorithms-discover how they work and implement from the scratch. Mach Learn Mastery 148:148–162
11. Brownlee J (2017) 00 ML Mastery—understand you data, create accurate models and work projects end-to-end. 感染症誌 91:399–404
12. Brownlee J (2019) Generative adversarial networks with python, deep learning generative models for image synthesis and image translation. Mach Learn Mastery 1–654
13. Brownlee J (2020) Discover how to harness uncertainty with Python. 319
14. Lee T, Singh VP, Cho KH (2021) Deep learning for time series. 107–131. https://doi.org/10.1007/978-3-030-64777-3_9
15. Al Shehri FH, Gryzlov A, Al Tayyar T, Arsalan M (2020) Utilizing machine learning methods to estimate flowing bottom-holepressure in unconventional gas condensate tight sand fractured wells in Saudi Arabia. Soc Pet Eng - SPE Russ Pet Technol Conf 2020, RPTC 2020. https://doi.org/10.2118/201939-ru
16. Sun J, Ma X, Kazi M (2018) Comparison of decline curve analysis DCA with recursive neural networks RNN for production forecast of multiple wells. SPE West Reg Meet Proc 2018-April: https://doi.org/10.2118/190104-ms
17. Zhang T, Song S, Li S, et al (2019) Research on gas concentration prediction models based on lstm multidimensional time series. Energies 12. https://doi.org/10.3390/en12010161
18. Yoo TW, Oh IS (2020) Time series forecasting of agricultural products' sales volumes based on seasonal long short-term memory. Appl Sci 10:1–15. https://doi.org/10.3390/app10228169
19. Hyndman RJ, Athanasopoulos G (2018) Forecasting: Principles and Practice. Princ Optim Des 504
20. Bianchi FM, Maiorino E, Kampffmeyer MC, et al (2017) An overview and comparative analysis of recurrent neural networks for short term load forecasting
21. Sahoo BB, Jha R, Singh A, Kumar D (2019) Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys 67:1471–1481. https://doi.org/10.1007/s11600-019-00330-1
22. Pandey YN, Rastogi A, Kainkaryam S, et al (2020) Machine learning in the oil and gas industry
23. Nachiketa Chakraborty (2000) Testing RNN-LSTM forecasting with simulated astronomical lightcurves. Preprints 1–7. https://doi.org/10.20944/preprints201907.0241.v1
24. Kinoshita M (2019) Capstone project : google stock prediction with deep learning models overview : statement :
25. Yoo TW, Oh IS (2020) Time series forecasting of agricultural products' sales volumes based on seasonal long short-term memory. Appl Sci 10:1–15. https://doi.org/10.3390/app10228169
26. Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. Princ Optim Des 504
27. Adam G, Josh P (2017) Deep learning: a practitioner's approach

Chapter 21
Predictive Analytics Model of an Engineering and Technology Campus Placement

Sachin Bhoite, Anuradha Kanade, Punam Nikam, and Deepali Sonawane

1 Introduction

According to the requirement of industry, colleges must update their curriculum and provide necessary technical and practical knowledge to the students. It will help in fulfilling the requirement of skilled and qualified students of the industries. DM and machine learning (ML) scholars have studied classification problems most recurrently [1]. In which the value of a dependent variable can be predicted based on the values of other independent variables [2]. This paper aims to determine the features impacting on prediction of placement and also students will get to know the placement status and get help in improving their weaker areas in advance.

Basically, this model will help to make training and placement officers (TPO) work easy and increment the total number of placements. Hence, it will directly lead to an increment in the rank of engineering and technology institutions. As our objective is to predict the placement of a student, in such a way that either he will get placement or not. It is a binary classification problem. To get good accuracy with minimum error, we have experimented with various classification ML algorithms with K-fold cross-validation techniques and trained and tested the data splitting techniques. The Value of K is tested for better results though most of the time it has

S. Bhoite (✉) · A. Kanade · P. Nikam · D. Sonawane
School of Computer Science, MIT-WPU, Pune, Maharashtra, India
e-mail: sachin.bhoite@mitwpu.edu.in

A. Kanade
e-mail: anuradha.kanade@mitwpu.edu.in

P. Nikam
e-mail: punam.nikam@mitwpu.edu.in

D. Sonawane
e-mail: deepali.sonawane@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_21

considered as 10. Also, we used EL techniques, which are comparatively faster and give better accuracy for classification projects.

2 Related Work

The researchers have studied several connected national and international research papers, thesis to understand datasets, data pre-processing methods, features selection methods, type of algorithms used in the existing studies.

Authors in [3] performed a step-wise analysis based on specific statistical frameworks for the placement. The analysis concluded with student datasets including academic and selection subtleties is important for forecasting future selection possibilities. Authors in [4] proposed the campus placement prediction work using the classification algorithms Decision Tree and Random Forest. The accuracy obtained after analysis for Random Forest is greater than the Decision tree. Authors in [5] used different ML algorithms to analyze students' admission preferences. They found Random Forest classifier is a good classifier as its accuracy is very high. Authors in [6] used different ML models to analyze students' placement, they found AdaBoost classifier along with the Bagging and Decision Tree as Base Classifier gives high accuracy. The student placement analyzer recommendation system, built using classification rules-Naïve Bayes, Fuzzy C Means techniques, to predict the placement status of the student to one of the five categories, viz., Dream Company, Core Company, Mass Recruiters, Not Eligible, and Not Interested in Placement. This model helps weaker students and provides extra care toward improving their performance henceforth [7]. Authors in [8] presented student career prediction using advanced ML techniques. In this paper, Advanced ML algorithms like SVM, Random Forest decision tree, One Hot Encoding, XG boost are used. Out of all, SVM gave more accuracy with 90.3%, and then the XG Boost with 88.33% accuracy.

Authors in [9] presented student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics. They used Support Vector Machine with RBF Kernel (SVM), Support Vector Machine with Linear Kernel (SVML), Logistic regression (LR), Decision tree (DT), Random Forest (RF) techniques. ML is used to predict placement results and the programming skill level. The researcher created a classification model with precision, recall, and F-measure.

Authors in [10] presented the study on educational data mining for student placement prediction using ML algorithms. ML algorithms are applied in the weka tool and R studio which are J48, Naïve Bayes, Random Forest, Random Tree, Multiple Linear Regression, binomial logistic regression, Recursive Partitioning, Regression Tree, conditional inference tree, Neural Network. In the weka tool, Random Forest and Random Tree algorithms are giving 100% accuracy on the student placement dataset. Authors in [11] presented a survey on placement prediction systems using

ML. The author has suggested ensemble methods, which is a Machine Learning technique that combines several base models in order to produce one optimal predictive model.

3 Research Methodology

The proposed work was carried out by performing experiments on the pass-out student's dataset with various ML algorithms.

3.1 Algorithms Used

The objective of research needs to use classification methods. Hence, researchers have used the following ML classification algorithms.

1. Logistic Regression
2. K-Nearest Neighbors
3. Decision Tree
4. Random Forest
5. Support Vector Machine
6. Naive Bays

   Also, used following advanced EL algorithms.

7. Adaptive Boosting,
8. Extreme Gradient Boosting (XGBoost) and
9. Grid Search CV

4 Steps in Building Predictive Models Using ML

We followed the Cross-Industry Standard Process (CRISP) methodology.

Understanding of problem and objectives of the research: Understanding dataset of already placed students and selection of the appropriate features for placement prediction.

Data Understanding: Data of already placed students were collected. All the attributes of the dataset were analyzed based on their importance and relevance based on the placement prediction. Point 5, About the dataset of this topic explains details about the dataset.

Feature Engineering: In this phase, the data from multiple data sources were integrated into one dataset. The next step is that the data were cleaned by removing unwanted columns, handling missing values, creating unique classes, performing transformation for numerical data, and all the cleaning activities on the data. Point 6, Feature engineering of this topic explains details about the same.

Table 1 Univariate feature selection for placement prediction

Feature name	Feature score	Feature name	Feature score
Sem_IV_Aggregate Marks	311.517737	Sem_VI_Pending Back Papers	22.920021
Aggregate Present Marks	223.533006	Sem_V_Pending Back Papers	20.066178
Sem_III_Aggregate Marks	198.255768	Sem_III_Back Papers	16.854853
Sem_VI_Aggregate Marks	151.002450	Back Papers	12.203902
Sem_V_Aggregate Marks	147.823502	Pending Back Papers	7.822886
Sem_II_Aggregate Marks	142.692722	Sem_I_Back Papers	5.458014
Sem_I_ Aggregate Marks	138.860021	Sem_II_Back Papers	2.265504
College Name	55.624319	Sem_II_Pending Back Papers	0.701740
Sem_VI_Back Papers	53.893101	Sem_IV_Pending Back Papers	0.558951
SSC Aggregate Marks	42.917186	Sem_III_Pending Back Papers	0.200665
Defense Type	27.490582	Sem_I_Pending Back Papers	0.124713
Category	27.385665	Gender	12.518480
12th/Diploma marks	26.351629	Branch	0.103342

Experimenting: A number of ML algorithms were tested and experimented with parameter tuning mentioned in Table 2 and 3 to predict the college, and its results are discussed in point 9, result and discussion.

Evaluation: Models developed were evaluated based on their performance for accuracy metric [13]. More information is presented in point 8.

Result and Discussion: Result and discussion are discussed in point 9.

Implementation: Once the model is evaluated, it is used to evaluate unseen data, which is discussed in point 10.

5 About the Dataset

Researchers have collected 16 engineering colleges’ 9766 data records. A number of columns in the dataset per college was varied from 20 to 46. We merged the dataset by considering common and important columns from our objective point of view in the excel file format, and then converted it into a CSV file. Which is essential to read by the python code to implement ML algorithms.

6 Feature Engineering

In general, every ML algorithm takes some input data to generate desired outputs. These input data are called features, which are usually presented in structured columns. As per goal or objective algorithms require input features with some specific

Table 2 List of experiments with model combinations

Sr No	Nameof Algorithm	Data splitting method used	Datasplitting folds/ratio			Parameter tuned	No. of parameter Tested
1	Logistic Regression	K-FCV	3	5	10	label encoding	6–10
						onehot encoding	6–10
		T-TS	70:30	80:20	90:10	label encoding	6–10
						onehot encoding	6 to 10
2	Support Vector Machine (SVC)	K-FCV	3	5	10	estimator	6–10
						param_grid	6–10
		T-TS	70:30	80:20	90:10	estimator	6–10
						param_grid	6–10
3	Decision Tree	K-FCV	3	5	10	max_depth	6–10
						min_impurity_decrease	6–10
						max_leaf_nodes	6–10
						min_leaf_nodes	6–10
						max_feature s	6–10
		T-TS	70:30	80:20	90:10	max_depth	6–10
						min_impurity_decrease	6–10
						max_leaf_nodes	6–10
						min_leaf_nodes	6–10
						max_feature s	6–10
4	Random Forest	K-FCV	3	5	10	max_depth	6–10
						min_impurity_decrease	6–10
						max_leaf_nodes	6–10
						min_leaf_nodes	6–10
						max_feature s	6–10
		T-TS	70:30	80:20	90:10	max_depth	6–10
						min_impurity_decrease	6–10

(continued)

Table 2 (continued)

Sr No	Nameof Algorithm	Data splitting method used	Datasplitting folds/ratio			Parameter tuned	No. of parameter Tested
						max_leaf_nodes	6–10
						min_leaf_nodes	6–10
						max_features	6–10
5	Gaussian NB	K-FCV	3	5	10		6–10
		T-TS	70:30	80:20	90:10		6–10
6	K Neighbors Classifier	K-FCV	3	5	10	leaf_size	6–10
						n_neighbors	6–10
		T-TS	70:30	80:20	90:10	leaf_size	6–10
						n_neighbors	6–10

Table 3 List of experiments with advanced algorithms

Sr. No	Name of the Algorithm	Data splitting method used
1	Ada Boost Classifier (DT)	T-TS
2	Extreme Gradient Boosting (XGBoost) Classifier	T-TS
3	Grid Search CV	T-TS

characteristic to get the desired output. Hence, there is a need of feature engineering. Feature engineering efforts mainly have two goals:

1. Generating the proper input dataset, as per the requirement of the ML algorithm.
2. Improving the performance of ML models.

As per the experience of the researcher, we need to spend more than 70% of the time on data preparation. The following steps are carried out to achieve the same.

1. Missing Values
2. Handling categorical data (Label Encoder)
3. Change the data type
4. Drop columns

7 Feature Selection

Every time domain experts may not be available to decide independent features to predict the category of the target feature. Hence, before fitting model, we must make sure that all the features that we have selected are contributing to the model properly and weights assigned to it are good enough so that our model gives satisfactory accuracy. For that, we have used 3 feature selection techniques: Univariate Selection,

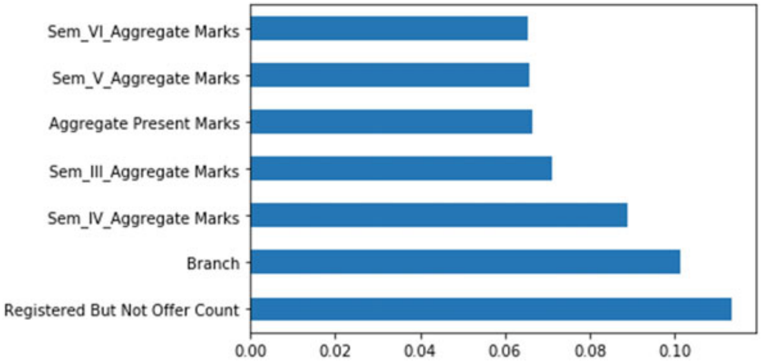


Fig. 1 Feature selection using feature importance for placement prediction

Recursive Features Importance, and Feature importance. We used the python scikit-learn library to implement it.

The Univariate Selection method shows the highest score for the following features (Table 1).

While using the Recursive Feature Importance method, the following features are selected, and the remaining are rejected.

Selected Features: [‘Pending Back Papers’, ‘Sem_III_Pending Back Papers’, ‘Sem_IV_Aggregate Marks’, ‘Sem_IV_Pending Back Papers’, ‘Sem_V_Pending Back Papers’, ‘Sem_VI_Back Papers’].

Inbuilt class Feature importance comes with Tree based Classifiers; we used Extra Tree Classifier from python scikit-learn library for extracting the top 7 features of the dataset (Fig. 1).

Hence, as per all the above methods and also as per domain our knowledge, we have chosen 25 important features which are as follows to predict target feature ‘Job Offer’.

[‘Branch’, ‘Aggregate Present Marks’, ‘Back Papers’, ‘Pending Back Papers’, ‘Sem_I_ Aggregate Marks’, ‘Sem_I_Back Papers’, ‘Sem_I_Pending Back Papers’, ‘Sem_II_Aggregate Marks’, ‘Sem_II_Back Papers’, ‘Sem_II_Pending Back Papers’, ‘Sem_III_Aggregate Marks’, ‘Sem_III_Back Papers’, ‘Sem_III_Pending Back Papers’, ‘Sem_IV_Aggregate Marks’, ‘Sem_IV_Back Papers’, ‘Sem_IV_Pending Back Papers’, ‘Sem_V_Aggregate Marks’, ‘Sem_V_Back Papers’, ‘Sem_V_Pending Back Papers’, ‘Sem_VI_Aggregate Marks’, ‘Sem_VI_Back Papers’, ‘Sem_VI_Pending Back Papers’, ‘12th/Diploma_Aggre_marks’, ‘SSC Aggregate Marks’].

8 Experimentation

There are adequate models that are studied and tested for the objective with optimal values for K-fold cross-validation (K-FCV), Train-Test split (T-TS), parameters tuning, and testing. In this process, the python sklearn library has played a very important role. So detail is mentioned in the table below.

Apart from the above methods while doing parameter tuning, we have used the following ensemble algorithms.

After the discussion of the accuracy results researcher has suggested a web module named 'Free guide to notify the campus placement status (FGNCPS)' through which students will get to know their placement status in advance and also come to know to work more on weaker areas.

9 Result and Discussion

After implementing data cleaning process, removing all the noise, selecting relevant features and encoded it into ML form, the next step is building a predictive model by applying various ML techniques to find out the best model which gives us more accuracy for train data and test data.

Model selection for placement prediction: After implementing all the above methods mentioned in Table 4 and 5, we found XGBoost classifier is the best classifier to predict campus placement.

Table 4 Results of placement prediction using ML techniques with K-fold cross validation

Sr. No	Name of Algorithm	Train Accuracy	Test Accuracy
1	Logistic Regression	0.7251336898395722	0.7371794871794872
2	Support Vector Machine	0.7235294117647059	0.7393162393162394
3	Decision Tree Classifier	0.8165775401069518	0.782051282051282
4	Random Forest Classifier	0.823663101604278	0.7162393162393162
5	Gaussian NB	0.5294117647058824	0.49145299145299143
6	K Neighbors Classifier	0.8235294117647058	0.7606837606837606

Table 5 Results of placement prediction using Ensemble Learning

Sr. No	Algorithm	Train Accuracy	Test Accuracy
1	AdaBoostClassifier(DT)	0.85	0.82
2	XGBoost	0.88	0.84
3	GridSearchCV	0.851336898395	0.82478632478632

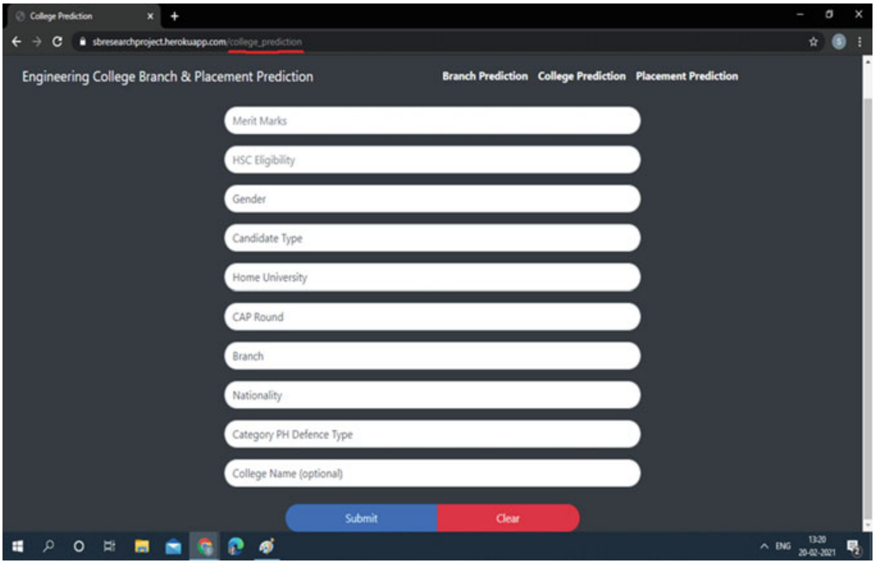


Fig. 2 Placement prediction web module

We can see the result of placement prediction using ensemble classifier XGBoost with 0.88 training accuracy and with 0.84 testing accuracy, which is comparatively very high. Hence, we have chosen the XGBoost classifier to implement the model.

10 Implementation

10.1 A Free Guide to Notify the Campus Placement Status (FGNCPS)

While predicting the campus placement of Engineering and Technology students, we have proposed the following FGNCPS web module. The aspirant student has to submit some basic information which is nothing but selected input features to predict their placement status in the early stage of academics (Fig. 2).

11 Conclusion

In this research, to predict the campus placement of Engineering and Technology students, all the ML model building steps are rigorously implemented on the dataset. Python, various libraries played a vital role during whole this process. In this study, 25 input features are selected out of the existing 46 features of the dataset. These features are very important, according to Univariate Selection, Recursive Features Importance, Lasso feature selection methods, and researchers' domain knowledge. To predict the campus placement, suit of ML and EL methods are experimented and compared. This suit contains Logistic Regression, K-Nearest Neighbors', Decision Tree Classifier, Random Forest Classifier, Naive Bayes, and Support Vector Machine classifiers. Under EL, we have experimented with Adaptive Boosting, Gradient Boosting, and GridSearchCV methods. After a comparison of all algorithms' accuracy, we found that the XGBoost classifier has greater accuracy for this project. Also, it has been observed that feature engineering is a very important step in a model building because, after it, results have been more improved. At the end researchers have suggested, 'A free guide to notify the campus placement status (FGNCPS)' web module for placement aspirant students.

References

1. Kabakchieva D, Stefanova K, Kisimov V (2011) Analyzing university data for determining student profiles and predicting performance. In: 4th International conference on educational data mining (EDM 2011). The Netherlands, pp 347–348
2. Nie M, Yang L, Sun J, Su H, Xia H, Lian D, Yan K (2017) Advanced forecasting of career choices for college students based on campus big data. Higher Education Press and Springer-Verlag Berlin Heidelberg
3. Kumar N, Singh AS, Thirunavukkarasu K, Rajesh E (2020) Campus placement predictive analysis using machine learning. In: 2nd International conference on advances in computing, communication control and networking (ICACCCN), ISBN: 978-1-7281-8337-4/20/$31.00 ©2020. IEEE
4. Pothuganti M, Swaroopa N (2019) Campus placement prediction using supervised machine learning techniques. Int J Appl Eng Res 14(9):2188–2191, ISSN 0973-4562
5. Kalathiya D, Padalkar R, Shah R, Bhoite S (2019) Engineering college admission preferences based on student performance. Int J Comput Appl Technol Res 8(09):379–384, ISSN:-2319–8656
6. Khandale S, Bhoite S (2019) Campus placement analyzer: using supervised machine learning algorithms. Int J Comput Appl Technol Res 8(09):379–384, ISSN:- 2319–8656, 358–362
7. Apoorva Rao R, Deeksha KC, Vishal Prajwal R, Vrushak K, Nandini (2018) Student placement analyzer: a recommendation system using machine learning. IJARIIE 4(3), ISSN(O)-2395-4396
8. Roy KS, Roopkanth K, Teja VU, Bhavana V, Priyanka J (2018) Student career prediction using advanced machine learning techniques. Int J Eng Technol 7:26–29
9. Ishizue R, Sakamoto K, Washizaki H, Fukazawa Y (2018) Student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics. Res Pract Technol Enhanced Learn 13. https://doi.org/10.1186/s41039-018-0075-y
10. Sreenivasa Rao K, Swapna N, Praveen Kumar P (2017) Educational data mining for student placement prediction using machine learning algorithms. Int J Eng Technol, [S.l.] 7(1.2):43–46, ISSN 2227-524X
11. Bangale M, Bavane S, Gunjal A, Dandhare R, Salunkhe SD (2019) A survey on placement prediction system using machine learning. IJSART 5(2), ISSN [ONLINE]: 2395-1052

Chapter 22
A Machine Learning Approach to Select Production Tubing Size for Oil Wells

Hrishikesh K. Chavan, Shubham T. Chavan, Saumya Koul, Shubham Kumar, Shailendra Naik, and Rajib Kumar Sinharay

1 Introduction

Production tubing selection is an important task in the life cycle of an oil well. It can directly impact the overall productivity of the well and plays a vital role in getting overall oil recovery factor. It is a tedious job, and as it changes, there is a change in productivity index [1]. Traditionally, nodal analysis is used to determine different production tubing sizes. This method consists of construction of inflow and vertical performance curves and choosing optimized tubing size using various correlations and graphical methods which would deliver maximum oil flow rate at minimum pressure drop [1–3]. This approach is found to be cumbersome and time consuming. It also leads to erroneous calculations of production tubing sizes, and there was a loss of productivity index [4–7]. Gilbert proposed methods that could utilize the whole curve for the determination of production tubing size. Gilbert considered five parameters that are wellhead pressure, flowing bottom hole pressure (FBHP), gas–liquid ratio (GLR), production rate and tubing depth [2]. A simulation and software approach is also given to select tubing size [5]. There is a gap between the actual production tubing size selected and production tubing size evaluated in order to accommodate several other parameters like casings, downhole equipment and future completion work. Moreover, gas liquid curves may not be available for each of the desired production rates. These problems can effectively be worked out by ML classification methods [8–10]. This paper tries to narrow this gap and incorporates results based purely on learning and training of the existing successful implementation production tubing sizes.

In this study, 239 production wells data are collected for four different suitable tubing sizes of 3.5”, 3.81”, 3.96” and 4” which are successfully used in various oil and gas field through literature survey. The data comprises parameters such as flowing

H. K. Chavan · S. T. Chavan (✉) · S. Koul · S. Kumar · S. Naik · R. K. Sinharay
Dr. Vishwanath Karad, MIT World Peace University, Pune 411038, India

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_22

bottom hole pressure (FBHP), gas–liquid ratio (GLR), production rate, degree API of crude oil, bottom hole temperature (BHT) and tubing head pressure (THP). The collected data is mined and cleaned using available techniques. This cleaned data is processed through different ML models such as perceptron model, support vector classification (SVC), Naive Bayes classifications, decision tree classifier, K-nearest neighbor classifier and random forest classifier (RF). The 70% production tubing sizes data is trained, and training accuracy is recorded. Remaining 30% is used for the testing, and testing accuracy is determined. The results are accurate and encouraging to determine the exact production tubing size for an oil or gas well.

2 Methodology

Following supervised learning models are used in this study.

2.1 Multilayer Perceptron Model

A multilayer perceptron model consists of input parameters, output neurons and hidden neurons. Input neurons are connected to the hidden layers, and hidden layers are connected to the output neurons. Activation function uses weights and bias to determine the output. This model is also called as backpropagation model because it modifies the weights and bias based upon error between actual value and calculated value in output neurons [11–13]. In this study, well and fluid parameters are used as input neurons, and tubing sizes are used as the output. A sigmoid function is used to determine accurate tubing size.

Sigmoid function for this perceptron model is given by Eq. 1:

$$\frac{1}{1+e^{-(\omega t(\text{input parameter})+b)}} \tag{1}$$

where

ω is weight factor, t is sigmoid variable, and b is bias constant.

Figure 1 shows the multilayer perceptron model used in this research work.

2.2 Random Forest Classification

This is a popular method of classification, and it is used before in the oil industry. Random forest classification randomly selects data, and it uses decision tree models to classify the data. It is an ensemble method in which data is divided into different sets. The sample data is chosen in a manner such that two data sets may or may

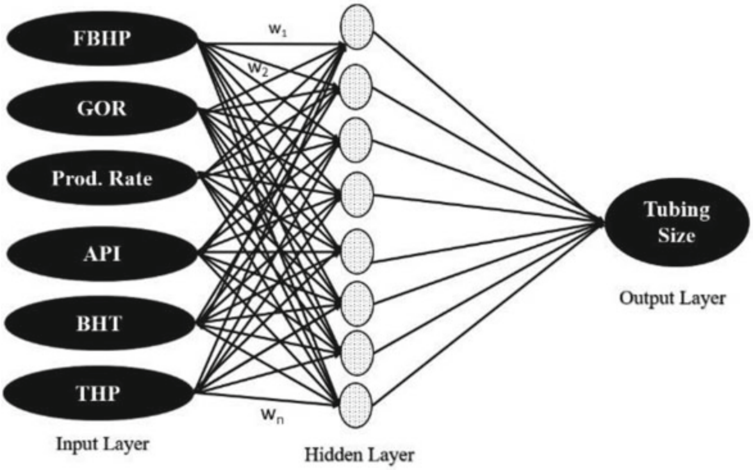


Fig. 1 Representation of perceptron model used in this study with input parameters

not have common data [14–16]. As shown in Fig. 2, the different models $M1$, $M2$, etc., are fed by different sample sets $S1$, $S2$, etc. Each model processes the data and produces output values. The repeated or the most occurring output value is selected at the end.

The final model $G(x)$ is given by following formula:

$$G(x) = M_1 + M_2 + \dots\dots\dots\dots\dots + M_n \tag{2}$$

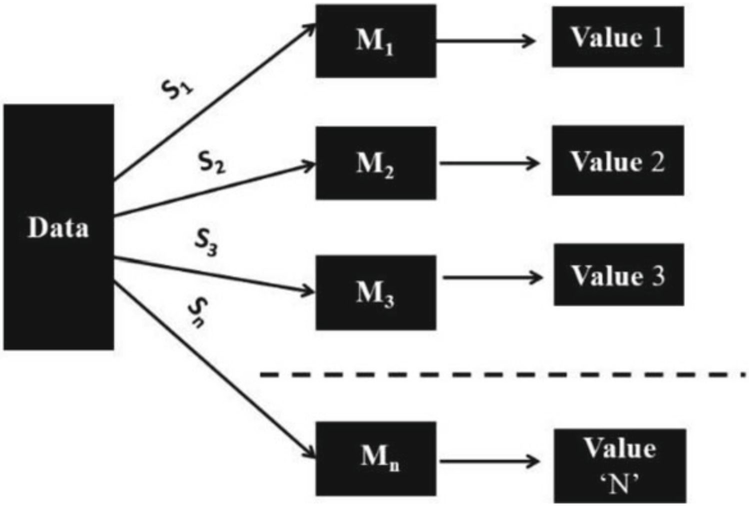


Fig. 2 Working of random forest classifier

where $G(x)$ is the final model. Sample sets S_1, S_2Sn are comprised of number of rows and columns.

2.3 Support Vector Classification

Support vector regression is characterized by the use of kernels and sparse solution. SVR problem formulation is best derived from a graphical perspective using one-dimensional example. A kernel helps to find hyperplane in the higher-dimensional space. The computational cost increases if the dimension of the data increases. This increases in dimension when it is impossible to find separating hyperplane in given dimension and requires higher dimension. Hyperplane basically separates two data classes in SVM [17, 18].

The function used to predict new values is

$$f(x) = \sum_{n=1}^{\infty} \left(\propto_n - \propto_n^*\right)\left(x_n' x\right) + b \tag{3}$$

where

$\propto$ is non-negative multipliers and x_n' is observations.

2.4 K-Nearest Neighbor (KNN)

K-nearest neighbor is a widely used algorithm because of its simplistic nature. It is a supervised learning technique which involves classification of labels based upon nearest value [19, 20]. This method calculates Euclidean distance which determines the closeness of the data points to the particular label. The formula for Euclidean distance is given by Eq. 4:

$$\text{Euclidean Distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{4}$$

where x_1, x_2, y_1 and y_2 are coordinates.

The shorter the Euclidean distance, there is great chance that it will be a part of that category.

2.5 Naive Bayes Classification

Naïve Bayes classification is one of the popular techniques which uses number of probabilistic algorithms together. This classification involves calculation of probability of occurring an event in the presence of another event [14, 21].

Bayes theorem uses following probability formula (Eq. 5) to predict the value:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{5}$$

where *y* represents tubing size and represents prediction of tubing size in the given conditions. *X* is a variable that represents features such as FBHP, BHT and THP.

3 Data Acquisition and PreProcessing

In this study, production tubing sizes of different oil wells were collected along with FBHP, THP, GLR, temperature, degree API of crude oil and production rate. The analysis of this data is shown in Table 1. The minimum flow rate obtained for this study was 280 stb/day, whereas minimum API of oil was 30. The maximum flowing bottom hole pressure was 6152 psi, and the maximum THP was 2600 psi. As shown in Table 1, the different statistical parameters were also calculated to understand the data.

The workflow as shown in Fig. 3 shows that noisy data has been removed before its processing. The data was then divided into training and testing set. Out of 239 data points, 70% data was used for training, and the remaining data was provided

Table 1 Statistics of the collected data

	FBHP (psi)	GLR (m3/m3)	Q (stb/day)	BHT	THP (psi)	API	SG
Maximum	6152	1.6	19,618	570	2600	38	0.876
Minimum	1227	0.03	280	88	88	30	0.834
Median	2541	0.53	4726	212	290	32.6	0.862
Arithmetic mean	2721.945	0.528	6211.336	217.294	380.563	33.905	0.855
Range	4925	1.57	19,338	482	2512	8	0.042
Standard Deviation	752.504	0.242	4886.15	27.349	277.549	2.269	0.011
Sample variance	566,262.338	0.058	23,874,463.02	3288.95	77,033.588	5.149	0.0001
Kurtosis	3.977	3.228	−0.322	15.925	24.516	−1.269	−1.231
Skewness	1.995	0.897	0.839	3.513	4.02	0.0209	0.01003

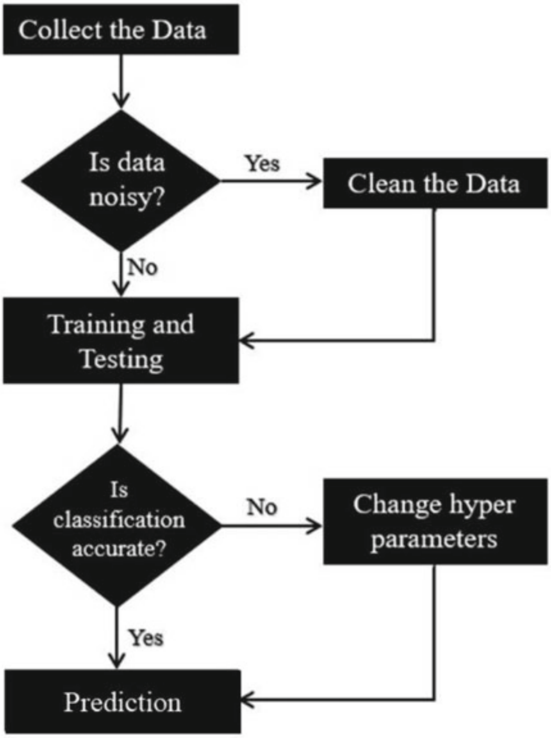


Fig. 3 Workflow adopted in this study

for the testing. Accuracy was calculated for the testing data. Hyperparameters were checked to attain maximum accuracy.

4 Result and Discussion

Heatmap was used to investigate the relationship between two input parameters. Heat plots can determine the visual relation effectively between two input variables. Figure 4 shows that there is positive relation between GLR and tubing head pressure (THP).

Flowing bottom hole pressure is correlated with THP. Other input parameters show flat relation between themselves (Fig. 4). Heatmap gives connection between input parameters. Heatmap is important because they give idea about how deeply input parameters are linked. Figure 4 shows the heatmap for this study. Heatmap study shows that flowing bottom hole pressure and THP are very well connected, whereas there is no link between THP and production rate. There is also good connection between production rate and API. It is important to understand the feature which has been doing strong impact on selection of tubing size. According to the study,

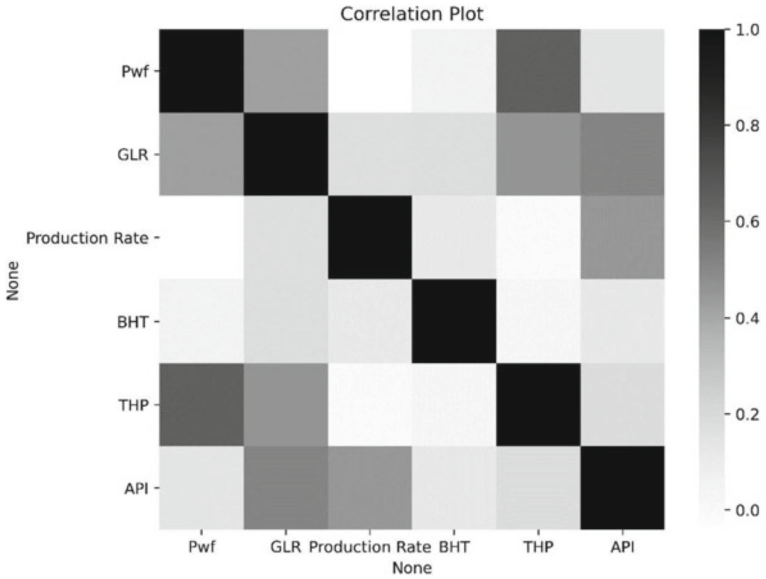


Fig. 4 Heatmap showing correlation between input parameters

THP is the most important parameter in the selection of tubing. Flowing bottom hole pressure and production rate also play vital role in choosing the production tubing. Study shows that API is not as important parameter as other parameters which can ultimately may affect in applying wrong tubing size.

Accuracy of all classification method was tested to identify best tool for tubing size selection. This study shows that random forest classification is the most accurate method to determine production tubing size. Other classification methods also show reasonable accuracy in getting the right tubing size (Table 2; Fig. 5).

Table 2 ML classification methods with accuracy score

ML methods	Accuracy
Random forest classifier	0.91
Support vector classifier	0.86
Perceptron model	0.86
Naive Bayes classifier	0.83
K-nearest neighbor	0.86
Decision tree classifier	0.86

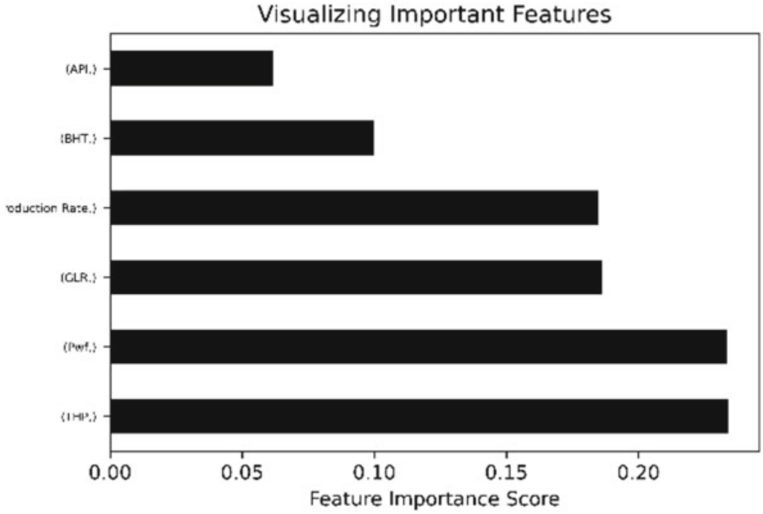


Fig. 5 Ranking of parameters affecting selection of tubing size

5 Conclusion

Machine learning can be a useful and unbiased technique in selecting the proper production tubing size. Random forest turns out to be the most accurate classifier to predict tubing size. According to the study, tubing head pressure is the most influential parameter while choosing the tubing size. Random forest can be used as an alternative tool to vertical lift performance to determine tubing size for an oil well.

Acknowledgements Authors would like to thank Dr. Vishwanath Karad MITWorld Peace University for allowing us to participate in the third Springer International Conference on Computational Science and Applications 2021 (ICCSA2021) which is organized by School of Computer Engineering, MIT-WPU. We would also like to extend thanks to School of Petroleum Engineering for the support.

References

1. Rempu W (2011) Selection and determination of tubing and production casing sizes. In: Advanced well completion engineering. Gulf Professional Publishing, pp 117–170
2. Guo B, Lyons W, Ghalambor A (2007) Petroleum production engineering. Elsevier Science & Technology Books
3. Beggs HD (1991) Production Optimization Using Nodal Analysis
4. Fevang O, Fossmark MG, Kulkarni KN, et al (2012) Vertical lift models substantiated by statfjord field data (SPE 154803). In: 74th European association geoscience engineering conference exhibition 2012 Inc SPE Eur 2012 responsibly security nature resource, pp 1772–1789

5. Nwanwe CC, Duru UI, Nwanwe OI et al (2020) Optimum tubing size prediction model for vertical multiphase flow during flow production period of oil wells. J Pet Explor Prod Technol 10:2989–3005. https://doi.org/10.1007/s13202-020-00964-8
6. Mogbolu E, Turan H, Rey-Fabret I, Okereke O (2014) Production forecast improvement using vertical lift performance curves: deep offshore Niger Delta. In: 38th Niger annual international conference exhibition NAICE 2014—Africa's Energy corridor oppor oil gas value maximization through integration glob approach, vol 2, pp1366–1378. https://doi.org/10.2118/172465-ms
7. Duns R (1963) Well control well performance 1 1. Heriot-Watt Univ 2:1–651
8. Suthaharan S (2016) Machine learning models and algorithms for big data classification
9. Aggarwal CC (2014) An introduction to data classification
10. Alkinani HH, Al-Hameedi ATT, Dunn-Norman S, et al (2019) Applications of artificial neural networks in the petroleum industry: a review. In: SPE middle east oil gas show conference MEOS, proceeding 2019-March. https://doi.org/10.2118/195072-ms
11. Kanin EA, Osiptsov AA, Vainshtein AL, Burnaev EV (2019) A predictive model for steady-state multiphase pipe flow: machine learning on lab data. J Pet Sci Eng 180:727–746. https://doi.org/10.1016/j.petrol.2019.05.055
12. Marius-Constantin P, Balas VE, Perescu-Popescu L, Mastorakis N (2009) Multilayer perceptron and neural networks. WSEAS Trans Circuits Syst 8:579–588
13. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408. https://doi.org/10.1037/h0042519
14. Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. ACM Int Conf Proceeding Ser 148:161–168. https://doi.org/10.1145/1143844.1143865
15. Wright MN, König IR (2019) Splitting on categorical predictors in random forests. PeerJ 2019:1–19. https://doi.org/10.7717/peerj.6339
16. Hegde C, Wallace S, Gray K (2015) Using trees, bagging, and random forests to predict rate of penetration during drilling. Soc Pet Eng—SPE Middle East Intell Oil Gas Conf Exhib. https://doi.org/10.2118/176792-ms
17. Brito J, Branco F (1991) Decision strategies for bridge management
18. Behzad M, Asghari K, Eazi M, Palhang M (2009) Generalization performance of support vector machines and neural networks in runoff modeling. Expert Syst Appl 36:7624–7629. https://doi.org/10.1016/j.eswa.2008.09.053
19. Devroye L, Wagner TJ (1982) 8 Nearest neighbor methods in discrimination. Handb Stat 2:193–197. https://doi.org/10.1016/S0169-7161(82)02011-2
20. Devroye+Wagner--1982--HandbookOfStatistics2-NearestNeighborMetyhodsInDiscrimination.pdf
21. Mücahid Mustafa Saritas AY (2019) Performance analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. Int J Intell Syst Appl Eng 3. https://doi.org/10.1039/b000000x

Chapter 23
Source Code Summarization & Categorization of Python and R Codes Using Deep Learning Technique

Shraddha Birari and Sukhada Bhingarkar

1 Introduction

Source code summarization is a methodology of understanding the programming language code and automatically generating the descriptions i.e., comments from the source code.

In software development life cycle i.e., implementation, testing, and maintenance most effortful phase is maintenance of the software system as 90% efforts are spent on maintenance [1]. Thus, to keep software system maintainable, documentation plays a vital role. According to the survey conducted in [2], 66% developers faced challenges while understanding motive or purpose of the piece of code. Understanding the code written by another person is a significant problem rated by 56% developers. 17% developers find it difficult to understand their own code written just a while ago. This clearly indicates the importance of well-written comments for the source code. Properly written comments are not only helpful while maintaining the software but also important for searching the relevant code. According to the survey conducted among developers regarding the code search [3], 33.5% people search to refer the example code. 26% people read or explore the code. 16% search for code localization like "*where class is instantiated*". 16% refer the code to determine the impact like "*understanding the dependencies*" or "*find side effect of proposed changes*"*0.8*.5% refer for metadata like "*who recently touched the code*". Also, according to the same survey, on an average developer creates 12 search queries as per each working day. This shows searching a code is a frequent and highly important activity. The performance of the code search is highly depending on the text involved in the search term and the code snippets [4]. Code search is difficult when search term

S. Birari (✉) · S. Bhingarkar
Computer Science and Engineering, MIT World Peace University, Pune, India

S. Bhingarkar
e-mail: sukhada.bhingarkar@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_23

specified as input do not have the same words as the corresponding source code. Thus, well-written comments will lead to effective code search.

In software development, finding relevant application in similar category is important to understand the functionality and useful for reusing the common functionality [5]. Programmers can save much time by finding the functionality in similar applications. For finding some application, one can search according to categories such as “Business”, “Communication”, “Audio/Video”, “Games” etc. [6]. The problem to find relevant functionality in similar category is due to the mismatch between descriptions written and implementation done.

Also, it is difficult for the newbie developer to understand and write proper comments and its relevant category. Writing the comments is a manual task; thus it can be a time-consuming work. Hence, it is important to have system with automatic comment generation with corresponding category.

Most of the existing comment generation techniques works on any single programming language. In this proposed work, we have used Python as well as R source codes to generate the comment with its corresponding category. LSTM based encoder and decoder along attention architecture is used to predict the comment and KNN algorithm predicts its corresponding category.

The rest of the paper is organized as follows: Sect. 2 elaborates related work about the source code summarization. Section 3 narrates the methodology of proposed source code summarization and classification. Section 4 gives experimental results. Finally, the paper concludes in Sect. 5.

2 Related Work

In this section, we present some existing techniques implemented so far for source code summarization.

In [7], Natural Language Processing (NLP) technique known as Statistical Machine Translation (SMT) is utilized which is mainly used for translation among two natural languages. SMT is rule based technique where rule is identified to convert source sentence into target sentence for given natural language like English to French etc. In this approach, Phrase Based Machine Translation (PBMT) is utilized that identifies phrase-to-phrase relationship among source and destination sentence. Consider $s = [s_1, s_2,..., s_{|s|}]$ as list of input sequence tokens, $t = [t_1, t_2,..., t_{|t|}]$ describes the array of output tokens. Collection of phrase pairs $\varphi = [\varphi_1, \varphi_2, ..., \varphi_{|\varphi}|]$ where $\varphi^n = |s^{(n)} \rightarrow t^{(n)}|$ which represents the sequence of the input sentence $s^{(n)}$ with its corresponding output subsequence $t^{(n)}$.The probability $Pr(t^{(n)}|s^{(n)})$ is calculated for the single phrase in the sentence by Phrase translation model. Then, $Pr(a|\varphi)$ known as reordering model calculates the probability for arrangement of every phrase. As code snippet has hierarchical structure which is not considered due to phrase-to-phrase translation in PBMT, and hence Tree-to-string Machine Translation (T2SMT) is introduced. In the T2SMT, instead of phrase pair φ “derivation” is introduced as d

= *[d_1, d_2,…, $d_{|d|}$]*. Each d^n shows the association among the input subtree $T_s^{(n)}$ and destination phrase $t^{(n)}$.

To maintain the hierarchical representation of the source code, [8] proposes Tree Based Convolutional Neural Network (TBCNN). In proposed work, Abstract Syntax Tree (AST) is used to present the code snippet. AST is constructed as node p with children c_1,.., c_n which is represented as a component $p \rightarrow c_1,\ldots,c_n$. Each node in the AST represented as a vector. Tree based convolutions used for feature detection which is calculated as

$$y = \tan h\left(\sum\nolimits_i^n W_{\mathrm{conv},i} \cdots x_i + b_{\mathrm{conv}}\right) \tag{1}$$

where n is number of nodes with vectors as $x_1,\ldots,x_n$, W and b are weight and bias of convolutional layer. As tree formed after convolution will be of same size as of input tree, thus dynamic pooling technique called one way pooling is used to handle variable size of tree.

CODE-NN model is proposed in [9], which uses LSTM model to predict the sentence for C# code snippets. In this approach, the attention mechanism calculates the distributional representation t_i of the code depending upon the current LSTM state h_i. Using t_i and h_i next word n_i is calculated. This CODE-NN approach used to retrieve the most appropriate source code from the corpus.

In [10], Natural Language Generation (NLG) based technique is proposed to generate the summary. In this approach, most important methods are identified using a PageRank algorithm. Then keywords are identified using the Software Usage Model (SWUM) which is basically used to identify the verbs, nouns, or prepositions from the statement. A custom NLG technique categorizes different types of messages according to their context. Then, according to the context of above messages, phrases are generated using lexialization. Finally, more readable phrases are generated in aggregation.

Reinforcement learning approach is implemented in [11]. Instead of considering only sequential tokens while representing the source code, this methodology uses AST based LSTM. Thus, structural as well sequential representation is considered in this work. For any random node of AST called j, hidden state h_{jl}, memory cell c_{jl} of lth child calculated as:

$$c_j = i_j \odot u_j + \sum_{l=1}^{N} f_{jl} \odot c_{jl} \tag{2}$$

$$h_j = o_j \odot \tan h(c_j) \tag{3}$$

Here ij is input gate,f_{jk} denotes the forget gate, o_j is output gate, u_j denotes a state to update the memory cell whereas the operator ⊙ denotes element wise multiplication between vectors. Next, the generated ASTs are converted to binary trees. For generating the summary, this work uses reinforcement learning based framework called

"Actor-Critic" network. The probability distribution of the next word concerning current state is calculated by actor. The model generates the *tth* word by following equation

$$p\pi\ (y_t|\,s_t) = softmax\,(W_s s_t + b_s) \tag{4}$$

where $p\pi\ (y_t|s_t)$ denotes the probability distribution of generating *tth* word y_t. S_t denotes the *tth* step. Finally, the critic network approximates the value generated at each *tth* step by using BLEU score to calculate the reward.

In [12], Source code Markup Language (SrcML) methodology is used for comment generation task. It translates the code written in Java, C and C++ to XML file [13]. Using input XML file, output document file is generated. It considers parameters, classes, conditions, and white spaces to generate the output XML File. Various tags such as <function> , <loop> , <class> considered where each tag in XML clearly show the main components of the source code. Feature extractor fetches the data generated from XML file. Using XPath, queries are performed to track each object from the code snippet. In this, important characteristics of the code will be extracted i.e., functions, attributes, calls, conditions. Attributes are basically the variables utilized in the given code, and it can be extracted using tag <decl> . Conditions are basically if conditions as well as several types of loops. Calls made by source code are categorized into calls, and it is in tag <call> and last is the <function> which is name of function and type along with the value returned by the functions. Then, the feature extractor generates file which contains program structure information. Finally, code descriptor generator reads two files one is source code, and another is program structure information file then generates the comments based on the source code and related information.

3 Source Code Summarization Methodologies

This section elaborates the proposed approach for source code summarization and classification on Python and R. In the proposed approach, LSTM networks [14] are used as encoder as well as decoder with Bahdanau attention mechanism [15] to generate the comments from the given source code. Also, KNN algorithm is utilized to classify whether the source code is of Python or R.

4 Source Code Summarization

The proposed source code summarization follows standard architecture of machine learning model. It involves encoder, decoder, and attention mechanism.

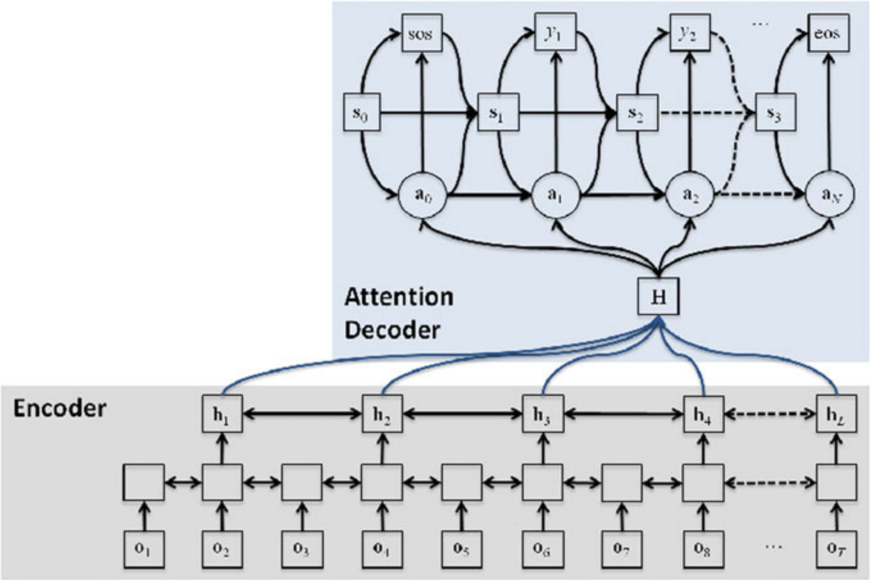


Fig. 1 Architecture of encoder-decoder model [16]

An encoder, in encoder-decoder is constructed to read input sequence, and then convert it to a predetermined length representation i.e., into encoded format. A decoder network uses output of encoder to generate output words until sequence token's end is reached. This architecture may not produce desired output due to variable size of input sequence and output sequence. By using the attention mechanism, model focuses on the tokens with a greater relevance.

Figure 1 shows the standard encoder-decoder architecture. The encoder transforms a source token O_i into feature sequence H, and then the decoder generates an output Y using the attention mechanism.

5 General architecture of LSTM cell

As mentioned earlier, in the proposed approach, LSTM networks are utilized to encode and decode the given sequence. LSTMs are the remarkable kind of Recurrent Neural Network (RNN) which can adopt the dependencies which are long termed. LSTM includes the gates which decides which information needs to be considered and which needs to discard. It has three important gates i.e., input gate, output gate as well as forgot gate which regulates the flow between LSTM cells. Forget gate determines the information that need to be kept or dropped using Sigmoid function. Current state's information and previous hidden state is passed to Sigmoid function which generates the value between 0 and 1. By the generated value, which is either

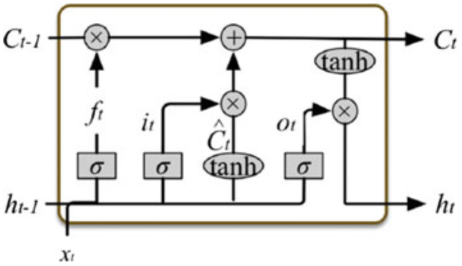


Fig. 2 Architecture of LSTM cell [17]

closer to 0 or 1 the information is discarded or kept, respectively. The cell state is updated using the input gate. At input gate, tan*h* function generates the value either −1 or 1 to pass the value to the network. Last, the output gate determines the upcoming hidden state. The product of sigmoid output and tan*h* output is used to generate the next state. Figure 2 shows the basic illustration of LSTM cell. In this architecture, $h_{t\text{-}1}$ shows the hidden state which is output from the preceding cell, and x_t is the input at given timestamp *t*. Below Eqs. (5), (6), and (7) give the calculations for f_t, i_t, o_t which represents the forget gate, input gate, and output gate, respectively. Here, w_f, w_i, w_o, b_f, $b_{i,}$ and b_o represent weights and bias for respective gates.

$$f_t = o'\left(w_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

$$i_t = o'\left(w_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

$$o_t = o'\left(w_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

Using above equations, final cell state denoted by c_t, candidate for all cell state ‘$c_{t,}$ and output of cell state h_t is calculated as below in Eq. (8), (9) and (10)

$$‘c_t = \tanh\left(w_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{8}$$

$$c_t = f_t * c_{t-1} + i_t * ‘c_t \tag{9}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{10}$$

6 Attention Architecture

Bahdanau attention is used to hold the relevant information. This attention mechanism also known as additive attention as it linearly combines encoder and decoder states. Encoder produces the hidden state h_j at each input token, using this hidden states and previous decoder's hidden state s_{i-1} alignment score e_{ij} i.e., alignment of input at placement *j* and output at placement *i* is calculated as follows

$$e_{ij} = a\left(s_{i-1}, h_j\right) \tag{11}$$

Alignment scores are passed to softmax activation to obtain the attention weights.

$$\alpha_{ij} = \exp\left(e_{ij}\right) / \sum_{k=1}^{Tn} \exp\left(e_{ik}\right. \tag{12}$$

Using above scores, context vector ci is calculated as

$$c_j = \sum_{j=1}^{Tx} a_{ij} h_j \tag{13}$$

This context vector is used by decoder to predict the target sequence.

7 Proposed source code summarization architecture

In proposed architecture, we have used python and R <code, comments> dataset as a training data. Initially, a tokenizer is prepared on the training data where source code snippet and corresponding comments are converted into tokens. Then, generated tokens are converted into corresponding integer sequence to proceed to the encoder-decoder model. Padding is done on the given integer sequence, where zero is padded up to the predefined maximum length i.e., 500.

We have used LSTM model to represents the code snippet and encode the given input code snippet. First layer is the embedding layer, which turns integer sequence into corresponding fixed length dense vectors. Embedding layer is constructed with the input size of the given code snippet vocabulary and 512 is given as latent space dimension i.e., dimension of the output embedding. First LSTM layer is designed with initial dimension of 512, which returns the output as hidden state h_i and cell state c_i and $i = 1$ at first layer. The above first LSTM layer's output is passed as input to the second layer and so it continuous up to the final encoder layer i.e., $i = 4$. The final LSTM layer generates the hidden state and cell state as *h* and *c*.

First layer at decoder is the embedding layer which is setup with the input size of comments or summary. Decoder LSTM is constructed with the above output of embedding layer, hidden state *h* and cell state *c* which is generated by the output of the final encoding layer.

Next, attention mechanism uses encoder states *h* as well as decoder states *s* combinedly to generate alignment score mentioned in Eq. (11). After calculating the alignment score of each encoder hidden state, these scores are combined and softmax activation is applied which in turn converts this alignment scores as a single vector as calculated in the Eq. (12). Finally, context vectors are calculated as referred in Eq. (13). The context vector concatenated with decoder's previous output and passed to decoder LSTM for given timestamp t to generate next output token. Finally, a dense layer with softmax activation generates the probability distribution over the target summary. The target integer sequence is decoded into its corresponding comment.

Figure 2 shows proposed architecture of source code summarization model which demonstrate generation of comment from given source code. As mentioned earlier, the given architecture is divided into three building blocks as mentioned earlier i.e., encoder, attention mechanism, and decoder. Consider the source code as:

def do_nothing():

return.

The above source code does nothing and has just a return statement. After preprocessing and tokenization above source code is converted into tokens as "def", "do_nothing", and "return". Initially, the above code snippet is converted into fixed length integer vectors using embedding layer. Later LSTM encoder generates hidden states *h* and cell state *c*. Decoder LSTM is configured with this state *h* and state *c* generated by encoder. The above encoder and decoder states are used by attention mechanism to generate the context vector c_i. This context vectors c_i with decoder's previous LSTM's output are used to generate the target sequence. This target sequence is converted into its corresponding comment. In this example, summary is generated as "returns nothing" (Fig. 3).

7.1 Source code classification

In proposed approach, we are generating the comments for Python as well as R. Thus, to classify whether the input source code is written in Python or R, KNN classification algorithm is implemented.

KNN is supervised technique which classifies the data upon the similarity among the data points. This algorithm starts by initializing the K value which is the value considered for the nearest neighbors. Calculate the distance with respect to considered K values using distance calculations like Euclidean distance. According to the distance calculated, assign the data point to its corresponding category.

In the proposed classification problem, we have considered the <code, category> pair where code is source code and category is its corresponding programming language i.e., Python and R. Initially, we converted the source code into vector of token count. The generated count matrix is converted into normalized Term Frequency–Inverse Document Frequency (tf-idf) representation. Generally, to show the importance of the word in given corpus tf-idf representation in used. Then we train the tf-idf representation of the code and category with its K parameter as 3. By considering the 3 as nearest neighbor value, it calculates Euclidean distance. By the

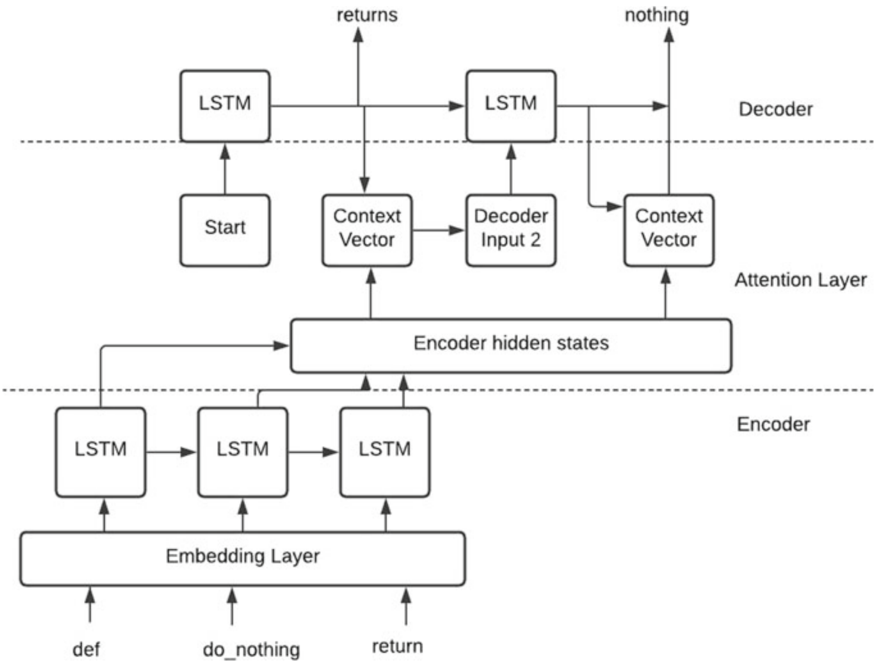


Fig. 3 Architecture of proposed source code summarization model

similarity score calculated, it determines the final category of the source code i.e., Python or R.

7.2 Dataset Preparation

For the summarization and classification problem, we have used dataset of 4000 <code, comment> pairs with its corresponding category as "Python" or "R". Python <code, comment> pairs are taken from dataset mentioned in [18]. For R, we have manually created the dataset with code and its corresponding comment.

7.3 Training Details

In proposed summarization model, we have used GPU to train the model. Model is trained on 50 epochs. The model is compiled using optimizer "RMSprop" and "sparse_categorical_crossentropy" as loss function.

Table 1 Evaluation score

Score name		Score
BLEU	BLEU-1	0.265
	BLEU-2	0.221
	BLEU-3	0.238
	BLEU-4	0.198
ROUGE-L	ROUGE-L (f)	0.274
	ROUGE-L (p)	0.275
	ROUGE-L (r)	0.287

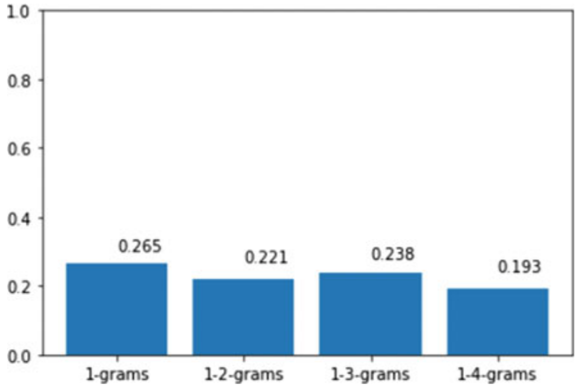


Fig. 4 BLEU evaluation

8 Evaluation

We have evaluated the proposed source code summarization model on the widely used evaluation criteria i.e., BLEU [19] and ROUGE-L [20].

BLEU shows how similar the predicted text is with respect to original text. Values closer to 1 represents the more similarity within the predicted output. ROUGE-L calculates the extensive matching series of words using Longest Common Subsequence (LCS).

Table 1 shows the above evaluation scores. BLEU 1, 2, 3, and 4 represent the 1-g, 2-g, 3-g, and 4-g whereas ROUGE-L f, p, and r represent the F1 score, Precision, and Recall, respectively.

Figures 4 and 5 show the above BLEU and ROUGE-L score, respectively.

9 Conclusion

Source code summary is useful to maintain the software system as well as to search the relevant code. In proposed paper, we have implemented source code summarization

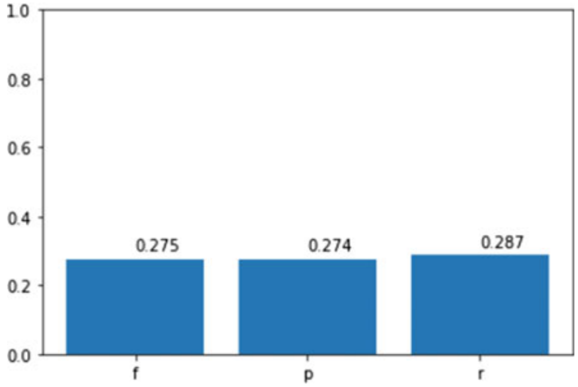


Fig. 5 ROUGE-L Evaluation

and classification methodology. We have used encoder-decoder which are LSTMs with attention to generate the summary. To classify the source code, we have used KNN algorithm. The proposed summarization technique is evaluated using standard evaluation criteria i.e., BLEU and ROUGE-L.

In the proposed summarization technique, we have used sequential content of the source code. In future, to preserve the hierarchical presentation of the source code snippet, we can implement LSTM that considers Tree structure of the source code.

References

1. Bennett KH (1990) The software maintenance of large software systems: management, methods and tools. In: BA Kitchenham (eds) Software engineering for large software systems
2. LaToza TD, Venolia G, DeLine R (2006) Maintaining mental models. In: Proceeding of the 28th international conference on Software engineering (ICSE '06)
3. Sadowski C, Stolee KT, Elbaum S (2015) How developers search for code: a case study. In: Proceedings of the 2015 10th joint meeting on foundations of software engineering
4. Nie L, Jiang H, Ren Z, Sun Z, Li X (2016) Query expansion based on crowd knowledge for code search. IEEE Trans Serv Comput 9(5):771–783
5. McMillan C, Grechanik M, Poshyvanyk D (2012) Detecting similar software applications. In: 34th International conference on software engineering (ICSE)
6. Nguyen AT, Nguyen TN (2017) Automatic categorization with deep neural network for open-source java projects. In: IEEE/ACM 39th International conference on software engineering companion (ICSE-C)
7. Oda Y, Fudaba H, Neubig G, Hata H, Sakti S, Toda T, Nakamura S (2015) Learning to generate pseudo-code from source code using statistical machine translation. In: 30th IEEE/ACM international conference on automated software engineering (ASE)
8. Mou L, Li G, Zhang L, Wang T, Jin Z (2016) Convolutional neural networks over tree structures for programming language processing. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
9. Iyer S, Konstas I, Cheung A, Zettlemoyer L (2016) Summarizing source code using a neural attention Model. ACL (1)
10. McBurney PW, McMillan C (2016) Automatic source code summarization of context for java methods. IEEE Trans Softw Eng 42(2):103–119

11. Wan Y, Zhao Z, Yang M, Xu G, Ying H, Wu J, Yu PS (2018) Improving automatic source code summarization via deep reinforcement learning. In: Proceedings of the 33rd ACM/IEEE international conference on automated software engineering
12. Mohsin AH, Hammad M (2019) A code summarization approach for object oriented programs. In: International conference on innovation and intelligence for informatics, computing, and technologies (3ICT)
13. Collard ML, Decker MJ, Maletic JI (2013) SrcML: an infrastructure for the exploration, analysis, and manipulation of source code: a tool demonstration. In: IEEE international conference on software maintenance
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
15. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
16. Ochiai T, Watanabe S, Hori T, Hershey JR (2017) Multichannel end-to-end speech recognition
17. Hu X, Li G, Xia X, Lo D, Jin Z (2018) Deep code comment generation. In:+ Proceedings of the 2017 26th IEEE/ACM international conference on program comprehension
18. Ahmad W, Chakraborty S, Ray B, Chang K-W (2020) A transformer-based approach for source code summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics
19. Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. 2002. s311–318
20. Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out

Chapter 24
An Analytical Survey on Heart Attack Prediction Techniques Based on Machine Learning and IoT

Nachiket Dunbray, Jayesh Katade, Sparsh Nimje, Shreyas Mavale, and Shamla Mantri

1 Introduction

A heart attack is a type of myocardial necrosis. This occurs because the coronary arteries that provide blood to the heart muscle might constrict due to an accumulation of fat, cholesterol, and other compounds are known as plaque. Atherosclerosis is the name for this gradual process. A blood clot develops around plaque within a cardiac artery when it breaks. This blood clot can obstruct blood flow to the heart muscle by blocking the artery. When the heart muscle is deprived of oxygen and nourishment, it develops ischemia. A heart attack, also known as a myocardial infarction, happens when ischemia causes damage or death to a portion of the heart muscle (MI).

Heart disease is a leading cause of death worldwide, and it has become a major health concern for many people. Cardiovascular Diseases (CVDs) claimed the lives of 17.9 million individuals worldwide in 2019, accounting for 32% of all deaths. Heart attacks and strokes were responsible for 85% of these deaths. CVDs were responsible for 38% of the 17 million premature deaths (before the age of 70) caused by noncommunicable diseases in 2019.

Many lives can be saved by the early discovery of the disease, and fatality can be minimized if patients receive treatment on time. These disorders are becoming more widespread even in younger age groups as a result of a lack of physical exercise caused by lifestyle changes. The primary causes of heart disease are consuming tobacco through cigarettes, shortage of physical activity, increase in cholesterol-rich foods, and poor living practices. Early detection of cardiovascular disorders such as heart attacks, coronary artery diseases, and other conditions is a key problem for regular clinical data analysis. Preventive medicine is becoming increasingly important and

N. Dunbray (✉) · J. Katade · S. Nimje · S. Mavale · S. Mantri
School of Computer Engineering and Technology, MIT World Peace University, Pune, India

S. Mantri
e-mail: shamla.mantri@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_24

popular around the world. In some cases, prevention is preferable to treatment. IHD events can be avoided in three ways: primordially, mainly, and secondarily [1]. Over the years, tons of data have been collected from various patients throughout the world. This data, although being collected, has never been put to good use. By using data mining, we can understand and use the insights found to further help improve research and prediction of the disease. We would then also be able to classify the patients of the low, medium, and high risk depending on the symptoms and medical history as well as the prediction outcomes from the various present techniques.

Machine learning is becoming increasingly popular in the medical diagnosis business, where computer analysis may reduce manual error and enhance accuracy. Machine learning algorithms make disease diagnosis more accurate. Machine learning techniques are used to forecast diseases such as heart disease, liver disease, diabetes, and tumors. Machine learning (ML) can help you make better decisions and make more accurate forecasts.

The provided dataset was best-fitted using RapidMiner and the processes of four algorithms were compared in RapidMiner which were: Random forest, K-Nearest neighbor, Naive Bayes, and Decision tree algorithm. The Support Vector Machine (SVM) was also used as a classification algorithm. According to surveys, almost 17 million people die each year as a result of cardiovascular disorders (CVD). Researchers in the field of machine learning frequently use this dataset.

Smart devices controlling and monitoring daily life have become a common occurrence in our society. IoT can be used to calculate the human factors that are needed to calculate the predictions. It can be used as direct contact with electronic devices over which the calculations are done. It is a convenient method to communicate with the human body because of its non-complexity and ease to store data. We can continuously monitor the human factors with the help of IoT which can help with the early prediction of heart attack and other diseases that can be prevented using preventive medicine.

2 Literature Survey

An intelligent Heart DIsease Forecast system was developed by Palaniap- pan and team [2] which predicts heart attack based on buried patterns and heart disorders [3] Use of distinct attributes in another model such as SVM, Logistic regression, and Random forest [2]

Bauder et al. used Oversampling and undersampling as two typical imbalanced data processing approaches which were researched in earlier studies [4]**.** Oversampling and undersampling cannot produce balanced features therefore the use of resampling and clustering techniques are discussed in the publication [5]. Many studies were based on two approaches for the prediction of a heart attack which were: based on medicine [6–8] and those which use patients' medical records using various artificial technologies [9]. Another method of random sampling was devised to select a set of samples and eliminate them [10]. In 2016 Fan et al and his team used an algorithm named as one-sided dynamic undersampling to decide which sample should be used in the classifier [11]. A K-means clustering undersampling algorithm for imbalanced data was employed by Shaik. Nakul and his team [12].

Various rules were created based on the PSO algorithm [13] and these rules were tested to find more accurate results. Following the evaluation of these rules, C.5.0 was used for classification using binary classifiers. In 2011 Soni et al. and his team used the decision tree algorithm, KNN algorithm, neural networks, for the prediction of cardiac disease [14].

3 Comparative Study of the Various Techniques

Upon research and comparison, we have found out that there are mainly 4 broad categories over which the prediction is done. We have briefly listed and discussed the categories below:

3.1 Classification

Classification is one of the widely used common data mining techniques. We split the data into categories of different groups based on their features and the data outcomes that we desire. We can define the goal of classification as to identify generic factors over which we can forecast unknown data to find their predictions. We can find a good explanation as to why the model did this particular decision based on the classifications that we have divided the data into.

3.2 Cluster

A cluster is a collection of items. For example, cluster partitioning the data set into cluster classes divides data items into multiple groups of similarity within a single group. Every nearby object qualifies as a neighborhood object. The cluster has two objectives. The first is an interclass competition, and the second is an intraclass competition. Cluster distance is maximum in an inter-class cluster. Cluster distances are kept to a minimum inside a cluster.

3.3 Feature Selection

This process of picking a subset of important features for use in model creation is known as feature selection, variable selection, attribute selection, and variable subset selection.

3.4 Association Rule

Data mining approaches rely heavily on association rule mining. The definition of an association rule is the identification of a large database of associations and their values. In this pattern, innovative tactics are used that do not help address classification or prediction difficulties. This work examines the challenge of departing a heart disease prediction task utilizing data mining approaches, as well as various issues in exiting a heart disease prediction task.

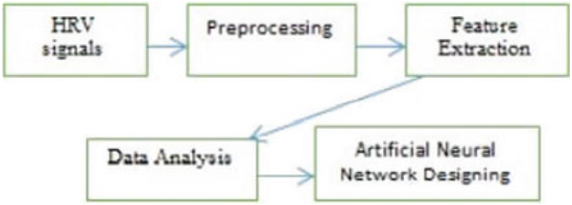
The broader aspect and techniques over which prior research has been done is discussed and evaluated in the following paper.

4 Vector Quantization Using Random Forest

Chen et al. devised a method for predicting cardiac disease. He used the Vector Quantization method, which is an artificial intelligence technique for categorization and prediction. It takes time to put data from earlier records to practical use. The accuracy rate is low. To combat this, the scientists utilized a random forest method to provide accurate findings in a shorter amount of time. The user inputs data, which is then compared to data already in the data collection using the Random Forest Algorithm. Random Forest is a classifier that uses more than one decision tree on various subsets of a dataset and averages the results to enhance the dataset's prediction accuracy. Instead of relying on a single decision tree, the RFA algorithm forecasts the outcome by combining the results of each decision tree (Figs. 1 and 2).

The initial step is to pick the K data points from the training set that has been chosen. Create as many decision trees as shown in Fig. 3 for the chosen data points. Decide on the number of decision trees you want to create. Reverse the first and

Fig. 1 System processing



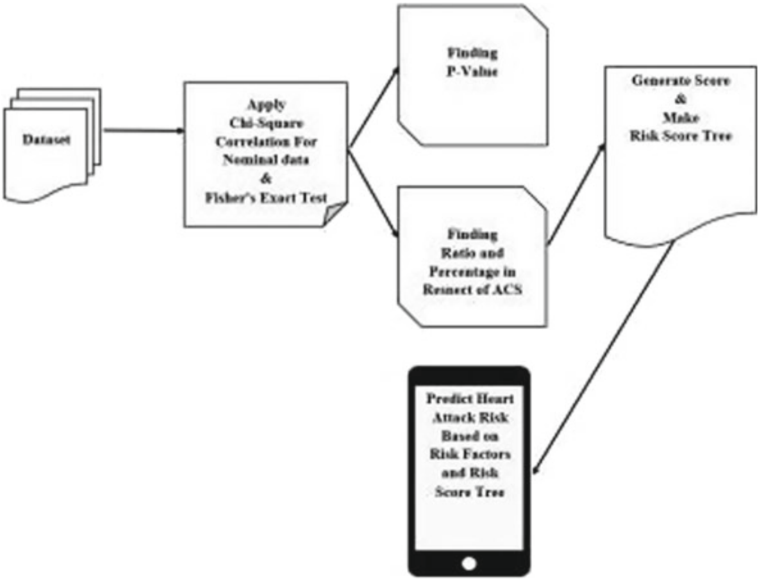

Dataset
Apply Chi-Square Correlation For Nominal data & Fisher's Exact Test
Finding P-Value
Finding Ratio and Percentage in Respect of ACS
Generate Score & Make Risk Score Tree
Predict Heart Attack Risk Based on Risk Factors and Risk Score Tree

Fig. 2 System architecture

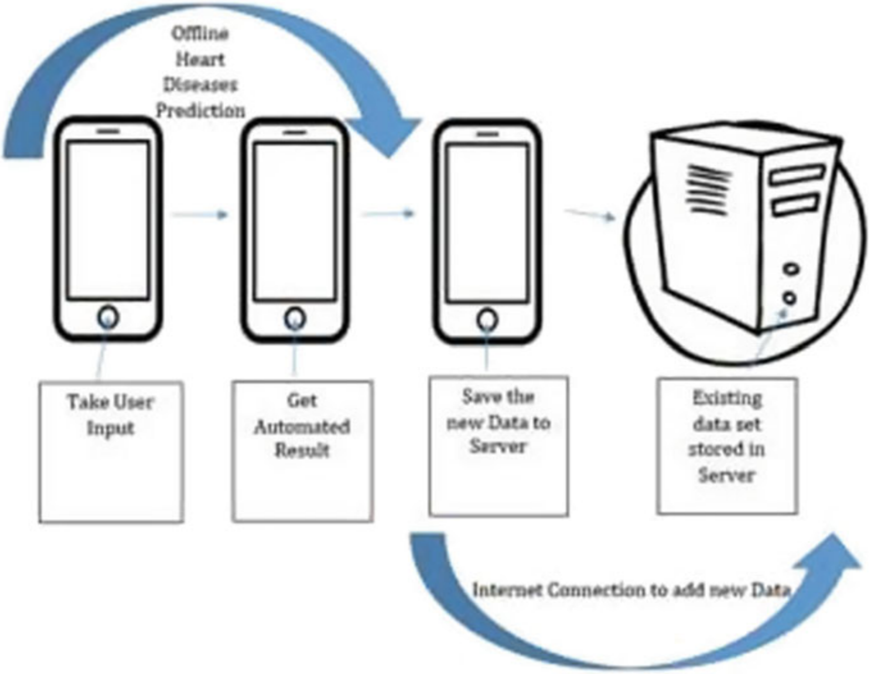

Offline Heart Diseases Prediction
Take User Input
Get Automated Result
Save the new Data to Server
Existing data set stored in Server
Internet Connection to add new Data

Fig. 3 Server communication with mobile app

second steps. Find the forecasts of each decision tree for new data points, and allocate the new data points to the category with the most votes. The application may be enhanced by adding new functions, such as sending a message to all of the user's family members if the user is diagnosed with heart disease. The information should also be sent to the closest hospital. Another option that should be accessible is an online medical consultation with the closest doctor.

5 Artificial Neural Network

The multiple forecasting range was chosen for prediction after evaluating the data acquired in all three periods. It has an efficiency of 94.28%. Furthermore, as per the way of treating VT, immediate therapy is administered within the time required for preparation (Fig. 3).

6 Chi Score Calculation

This team examined and linked clinical data from 787 individuals with risk factors such as hypertension, diabetes, dyslipidemia, smoking, family history, obesity, stress, and current clinical symptoms that might indicate underlying IHD. To integrate the clinical data collected from patients, they utilized an android application. A score was calculated after the data was mined using data mining technologies. In the case of IHD, the risks are divided into three categories: low, medium, and high. The score was used in the creation of an Android app.

Figure above depicts the system architecture that the team has used in their system. It shows how the data has been integrated and mined to understand the data and predict the heart attack risk.

Figure shows the mobile application communicating with the system servers. It shows how the user provides input and gets the results based on prior data.

The system consists of the following:

(a) Risk Score Tree: They utilized the chi-square correlation as well as the p-value to create this tree. The p-value was utilized to determine the significance threshold [15, 16]. A rating was assigned that was examined using ratios and percentages in the context of Acute Coronary Syndrome and a histogram were created (Fig. 2).

(b) Chi-Square Correlation: The nominal information correlation between two qualities is discovered using the Chi-square test [17–20]. The chi-square value is calculated using the formula [17–20]:

$$\tilde{\chi}^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

(c) p-value: The symbol p stands for probability. When the null hypothesis of the research is true, the p-value represents the likelihood of discovering the observed, or more extreme findings [21, 22].

(d) Fisher's Exact Test: It's for one contingency table only [22, 23].

(e) The Score Calculation: Age was classified into four types, and assigning weights depending on the amount of relevance and connection with IHD. Score 1 is for users between 15 and 30 years old, score 2 is for users between 30 and 40 years old, score 3 is for those between 40 and 55 years old, and score 4 is for users over 55 years old.

(f) Android Application Features: The produced data and score tree were incorporated into the mobile application. The user's details like their name, address, age, gender, occupation, contact information, and medical history of the Family as well as the user's medical history, drug history, and symptoms were all created. The user provides precise information regarding known risk factors and symptoms, based on which the program generates a rating and assigns the individual to a category (Fig. 3).

Figure 4 shows the application system of the model. They developed a working android application that uses the calculation and sends the results to the user based on the input data given.

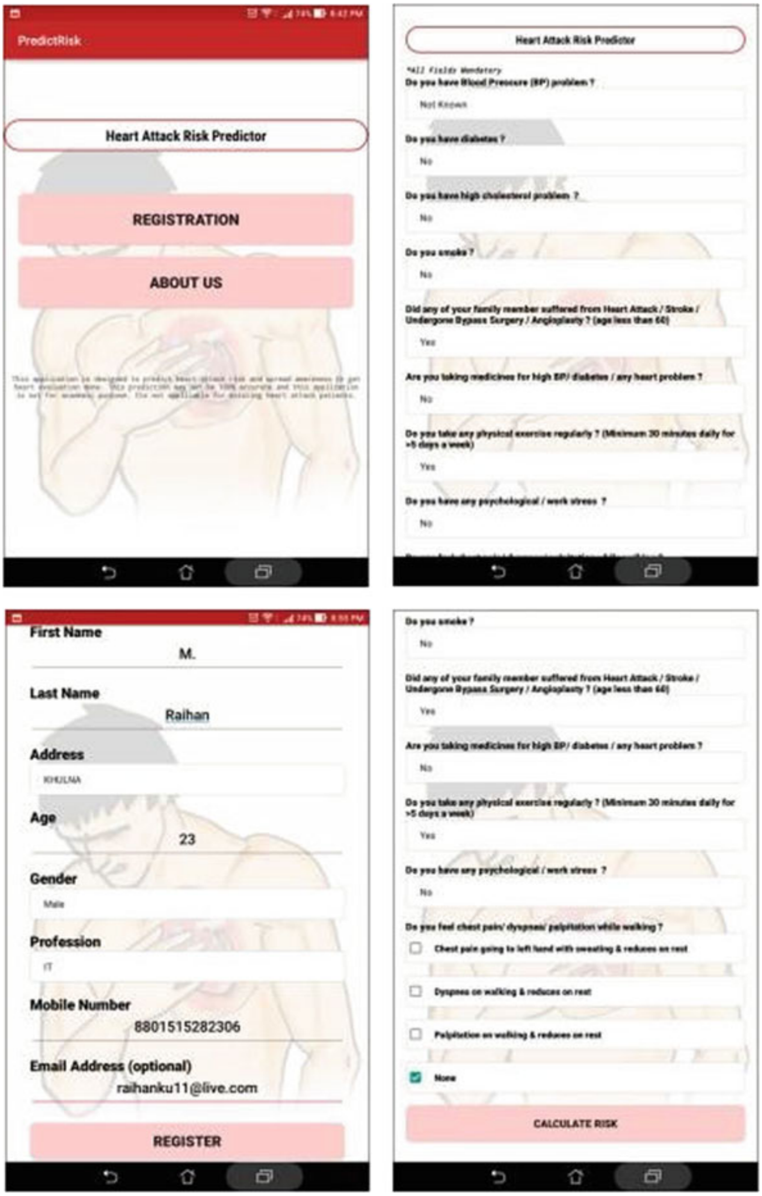
7 Internet of Things (IoT)

This block diagram shows the basic model depiction that was followed during the implementation of their system (Fig. 5).

They used a PPG sensor to detect the heart pulse. Photoplethysmography is the functioning concept of a heart rate sensor (PPG). Each heartbeat results in a change in blood volume when blood travels through an organ or a blood artery. A light is shone through the exposed skin, causing an intensity shift in proportion to the change in blood volume [24].

When the observed pulse rate is graphed, we obtain spectra with several spikes. This allows for the calculation of a pulse transit time (PTT) to respiration rate. Both the ECG's R wave and the PPG's peak can be recorded as a time or amplitude series. They may now be used to generate the PTT series. A curve-fitting approach is employed to soften the obtained data. These interpolated impulses are now subjected to an FIR filter. In most cases, the respiratory rate is computed in one-minute data segments.

The above graph shows the pulse received by a PPG sensor from a test subject. It shows the peaks of the pulse which can be used to further understand the patient's heart rate.



This application is designed to predict heart attack risk and spread awareness to get heart evaluation done. This prediction may not be 100% accurate and this application is not for academic purpose. Its not applicable for existing heart attack patients.

*All Fields Mandatory

Do you have Blood Pressure (BP) problem ?
Not Known

Do you have diabetes ?
No

Do you have high cholesterol problem ?
No

Do you smoke ?
No

Did any of your family member suffered from Heart Attack / Stroke / Undergone Bypass Surgery / Angioplasty ? (age less than 60)
Yes

Are you taking medicines for high BP/ diabetes / any heart problem ?
No

Do you take any physical exercise regularly ? (Minimum 30 minutes daily for >5 days a week)
Yes

Do you have any psychological / work stress ?
No

First Name
M.

Last Name
Raihan

Address
KHULNA

Age
23

Gender
Male

Profession
IT

Mobile Number
8801515282306

Email Address (optional)
raihanku11@live.com

Do you smoke ?
No

Did any of your family member suffered from Heart Attack / Stroke / Undergone Bypass Surgery / Angioplasty ? (age less than 60)
Yes

Are you taking medicines for high BP/ diabetes / any heart problem ?
No

Do you take any physical exercise regularly ? (Minimum 30 minutes daily for >5 days a week)
Yes

Do you have any psychological / work stress ?
No

Do you feel chest pain/ dyspnea/ palpitation while walking ?

□ Chest pain going to left hand with sweating & reduces on rest

□ Dyspnea on walking & reduces on rest

□ Palpitation on walking & reduces on rest

☑ None

Fig. 4 Android application

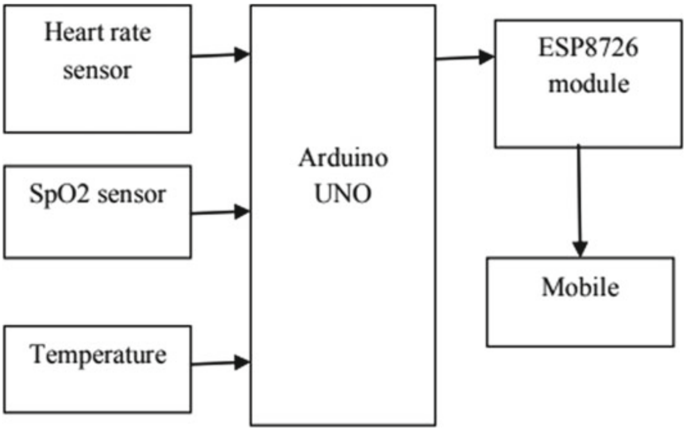


Fig. 5 Block diagram

They used a SpO2 sensor in their system as well. The amount of oxygenated blood is measured by SpO2, a biological metric. It may be characterized as the measure of oxygenated hemoglobin in red blood cells compared to non-oxygenated hemoglobin (Fig. 6).

The variation in luminosity can be used to detect SpO2 using a transmissive setup of a sensor and detector. The color observed in the detector varies when the intensity of red changes, and from this, the spO2 value may be calculated. For a healthy person, normal SpO2 levels range from 95 to 100% [25].

The system consisted of a temperature sensor. Because the circulatory processes of the human body play a crucial role in body thermoregulation, temperature as a vital indicator plays a significant influence on an individual's heart health [24–27] (Fig. 7).

The table above shows the evaluation and comparison of the different test subjects and observed levels of different factors that are being used to calculate the prediction of a heart attack.

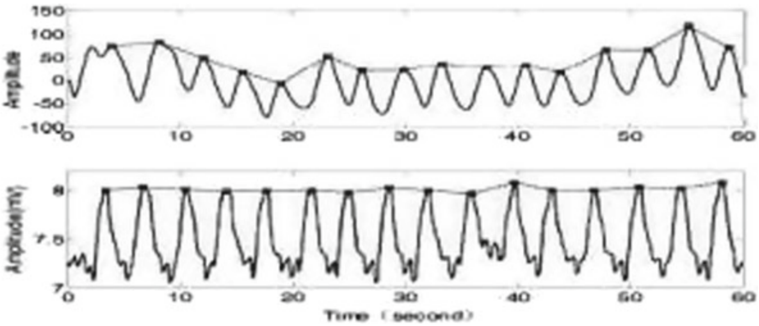
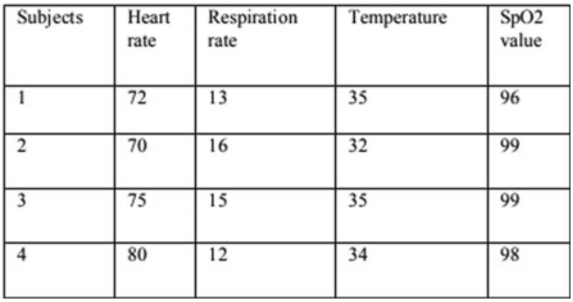


Fig. 6 Peak to peak plot of PPG



Subjects	Heart rate	Respiration rate	Temperature	SpO2 value
1	72	13	35	96
2	70	16	32	99
3	75	15	35	99
4	80	12	34	98

Fig. 7 Test subjects' comparative analysis of factors

8 Under Sampling—Clustering—Oversampling (UCO) Algorithm

The UCO method was created to handle unbalanced data from stroke patients. The UCO method can analyze the dataset and produce an approximately balanced Training dataset.

Using resampling and clustering approaches, Unbalanced data is processed using a specified algorithm. Under sampling and oversampling are two types of resampling. An under sampling method was used to remove the insignificant data samples.

To expand the size of minority samples, oversampling is performed. Clustering is the process of dividing data into groups and ensuring that samples in each category have similar characteristics. Clustering is performed on the minority label sample set.

The sample-set obtained by the under sampling process, as well as each cluster formed, were split into a training set, validation set, and testing set in the 8:1:1 ratio. The training sets that have been achieved are combined. The validation and testing sets that have been obtained are as well. The SMOTE method is then used to produce the final training set by oversampling the samples in the combined training set with minority labels. Every type of positive sample is represented in the final training set, and the number of negative samples is equal to the number of positive samples.

Machine learning algorithms can learn how to extract features from a training set like this. Finally, the technique creates a nearly balanced training set, validation set, and testing set. The suggested method is known as the Under sampling-Clustering-Oversampling (UCO) algorithm because of its concise name.

The UCO algorithm is made up of three components. Because the number of individuals who suffer a heart attack is far smaller than the number of people who do not, the samples of people who haven't had a heart attack are under sampled. Random under sampling yielded a total of 120 samples. During training, a clustering operation with a heart attack is done over the final samples to get the needed characteristics of each set of close-distance data. Finally, during the training phase, the samples with heart attacks are oversampled so that the number of heart attack samples is almost equivalent to the number of heart attack samples.

Figure 7 illustrates the workflow of the techniques being used. All the data collected about the stroke patient is analyzed in the first step. Then a data processing algorithm is designed as discussed above. Two analyses of the algorithm are done based on its complexity and its module. Finally, various comparisons of results are done.

9 Binary Classification Model

Binary classification is a classification process or task in which a set of data is divided into two groups. It’s essentially a guess as to which of two categories the object belongs to (Fig. 8).

An individual’s heart attack risk factor was estimated using his or her medical records and providing the data to the system using binary classification. The designed system has a simple and comprehensive user interface. The categorization is based on supervised learning, with the dataset coming from the machine learning repository at the California State University, Irvine.

A total of 14 attributes were present in the dataset which included 13 predictor variables and one binary variable. The classification method utilized was Naive Bayes.

The features of the dataset which will be useful for the prediction of heart attack are as follows: Age, Sex, Peak heart rate achieved, Chest pain, Electrocardiographic

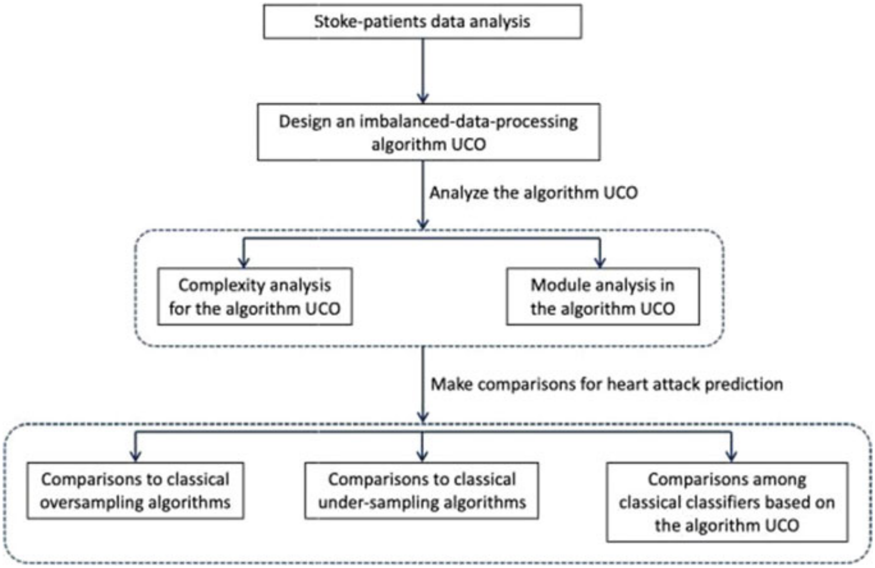

Stoke-patients data analysis
Design an imbalanced-data-processing algorithm UCO
Analyze the algorithm UCO
Complexity analysis for the algorithm UCO
Module analysis in the algorithm UCO
Make comparisons for heart attack prediction
Comparisons to classical oversampling algorithms
Comparisons to classical under-sampling algorithms
Comparisons among classical classifiers based on the algorithm UCO

Fig. 8 Flow diagram of UCO algorithm

results at Rest, Cholesterol, Blood Pressure at Rest, Fasting blood sugar, Slope of the peak exercise ST segment, “Num”—response variable, Exercise-induced angina, Number of major vessels, ST depression induced by exercise relative to rest, Thal.

For classification, the SciKit-learn package’s Gaussian Naive Bayes method was used. Scikit-learn is a collection of data-mining applications. Training and test data were separated from the processed dataset. The training data was transformed to a NumPy array before using the Gaussian Naive Bayes method. Once the test data was fed into the regression, the accuracy and other parameters were computed. A simple and interactive interface was built to assess the risk factor. A Python-based CGI script was used which accepts the medical records as input data to the classification algorithm, which in return calculates the risk factor for each individual.

10 Hybrid Model (DT + RFA)

The probabilities acquired from one machine learning model are fed into another machine learning model as input in a hybrid model, which is a novel approach. For the implementation of this hybrid model, various libraries are essential such as Pandas, sklearn libraries, MatPlotLib, etc.

The hybrid model was created over the following steps. First, the data was collected from the UCI data repository. EDA was done, therefore, after which the data was split into training and testing data categories. They had initially tested the data on the decision tree and random forest as separate models but the results were not satisfactory. To overcome this they decided to build a hybrid model based on the two of them as an ensemble model which increased the accuracy of the model. Finally inputting the data from a user’s perspective they got the prediction from the model.

We develop a hybrid model using a decision tree algorithm and a random forest algorithm. Random forest assumptions are used in the simulation. The training data is merged with the random forest statistics and sent into the decision tree classifier. In a manner analogous, decision tree probabilities are found and fed to gather the required information. In the end, values are expected. The anticipated cardiac illness for the testing dataset is presented when machine learning has been applied to a preprocessed dataset. The user/patient can enter their data to assess their cardiovascular disease risk. A categorical predictive category is used to classify the disease, with 0 denoting normal and 1 denoting cardiovascular disease.

11 Conclusion

Heart attack is one of the world’s most prominent chronic diseases that turn out to be fatal in most cases. These disorders are becoming more widespread even in younger age groups as a result of a lack of physical exercise caused by lifestyle changes.

The primary causes of heart disease are high cholesterol foods, poor living practices, smoking, junk food, and lack of physical activity. Early detection of cardiovascular disorders such as heart attacks, coronary artery diseases, and other conditions is a key problem for regular clinical data analysis. Preventive medicine is becoming increasingly important and popular around the world. In some cases, prevention is preferable to treatment. The above-studied techniques help to prove how we can find and detect heart attack at an early stage and even before the occurrence. Doctors can then prescribe preventive medications that will help in the survival of the patients. After analyzing various techniques to detect heart attacks, we conclude that most of the techniques can successfully detect heart attacks at a preventive stage.

References

1. Coronary Heart Disease in Bangladesh (2014) World Life Expectancy, 2014. [Online]. Available http://www.worldlifeexpectancy.com/bangladeshcoronary-heart-disease
2. Palaniappan S, Awang R (2008) Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS international conference on computer systems and applications, Doha, pp 108–115. https://doi.org/10.1109/AICCSA.2008.4493524
3. Hertzong M, Poezehl B (2010) Cluster analysis of symptom occurrence to identify subgroups of heart failure patients: a pilot study. J Cardiovasc Nurs 25:273–283
4. Bauder RA, Khoshgoftaar TM, Hasanin T (2018) Data sampling approaches with severely imbalanced big data for medicare fraud detection. In: Proceedings IEEE 30th international conference tools artificial intelligence (ICTAI), November 2018, pp 137–142
5. Wang M, Yao X, Chen Y (2021) An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients. IEEE Access 9:25394–25404. https://doi.org/10.1109/ACCESS.2021.3057693
6. Fenglan Z (2001) Electrocardiogram QTcd changes and prognosis in different periods of acute myocardial infarction. Shanxi, China: Shanxi Clinical Medicine
7. Yuejin Y et al. (2020) Spontaneous improvement of exercise abnormality in acute myocardial infarction and the predictive value of low-dose dobutamine echocardiographic test. China J Circulat., to be published
8. Jianping Q, Hong H (2013) Analysis of the predictive value of high frequency electrocardiogram on acute myocardial infarction. Biomed Eng Res to be published
9. Yaowang L et al. (2020) Application of machine learning algorithm in prediction of coronary heart disease and myocardial infarction. Int Med Health Guidance News, to be published
10. Little MA (2019) Random sampling. In: Machine learning for signal processing
11. Fan Q, Wang Z, Gao D (2016) One-sided dynamic undersampling nopropagation neural networks for imbalance problem. Oxford, U.K., Pergamon Press
12. Nagul S, Kumar RK (2021) An effective K-means approach for imbalance data clustering using precise reduction sampling. Int J Comput Sci
13. Azhar Hussein A et al. (2017) Using PSO algorithm for producing best rules in diagnosis of heart disease. In: 2017 international conference on computer and applications (ICCA). IEEE
14. Jyoti S et al (2011) Predictive data mining for medical diagnosis: an overview of heart disease prediction. Int J Comput Appl 17.8:43–48
15. Ahmed K, Jesmin T, Zamilur Rahman M (2013) Early prevention and detection of skin cancer risk using data mining. Int J Comput Appl 62(4):1–6
16. Jesmin T, Ahmed K, Zamilur Rahman M, Badrul Alam Miah M (2016) Brain cancer risk prediction tool using data mining. Int J Comput Appl 61
17. Witten I, Frank E, Hall M (2011) Data mining. Morgan Kaufmann, Burlington, MA

18. Han J, Kamber M, Pei J (2012) Data mining. Elsevier/Morgan Kaufmann, Amsterdam
19. Tan P, Steinbach M, Kumar V (2005) Introduction to data mining. Pearson Addison Wesley, Boston
20. “Statistical Analysis 5: Chi–squared test for 2–way tables”, statstutor. [Online]. Available http://www.statstutor.ac.uk/resources/uploaded/coventrychisquared.pdf
21. “P Values” (2016) Statsdirect.com [Online]. Available http://www.statsdirect.com/help/default.htm#basics/pvalues.htm
22. “Fisher’s exact test”, Wikipedia. [Online]. Available https://en.wikipedia.org/wiki/Fisher%27sexacttest#citenote-1
23. Mehta CR, Patel NR (2012) IBM SPSS exact tests. University of Sussex, 2012. [Online]. Available http://www.sussex.ac.uk/its/pdfs/SPSSExactTests21.pdf
24. Coyle S, Morris D, Lau K-T, Diamond D, Moyna N (2009) Textile-based wearable sensors for assisting sports performance. IEEE https://doi.org/10.1109/P3644.56
25. Sidheeque A, Kumar A, Balamurugan R, Deepak KC, Sathish K (2017) Heartbeat sensing and heart attack detection using internet of things: IoT . Int J Eng Sci Comput
26. Teichmann D, Kuhn A, Leonhardt S, Walter M (2014) The MAIN shirt: a textile-integrated magnetic induction sensor array. Sensors 14:1039–1056. https://doi.org/10.3390/s140101039
27. Rotariu C, Cristea C, Arotaritei D, Bozomitu RG, Pasarica A (2016) Continuous respiratory monitoring device for detection of sleep apnea episodes. In: 2016 IEEE 22nd international symposium for design and technology in electronic packaging (SIITME)

Chapter 25
IoT-Based Intelligent Inhaler System with Temperature, Oximetry Measurement, and Emergency Functionality

Vrushabh Sachin Shah and Vinayak Musale

1 Introduction

With ever-increasing pollution, the number of respiratory conditions has also increased, which is not only introducing people to a multiplicity of respiratory disorders but making it difficult for people with asthma and COPD to live a normal life. It is difficult to control the pollution, but we can still prevent the damage from it if we keep monitoring the pollution levels and take preventive measures. Some of the typical problems with asthma patients include indecisiveness about when to consult a doctor, the angst of going out alone, being unable to locate the inhaler during an attack, dubious about the current medical situation, and so on.

The smart inhaler is a preferment over the original inhaler in the sense that it does everything the traditional inhaler can and much more. It can keep track of the usage in terms of weeks and months, which benefits in identifying a medicine that suits the patient while surveilling the air quality incessantly and forewarn the patient about the degraded air quality. The proposed system facilitates the patients to measure, store, and process various data like body temperature, heart rate, and Spo2, which can then be amalgamated with even more data, into meaningful medical reports. These medical reports generated can be stored on the mobile phone and can be shared with any of the doctors available on the doctor's network, replacing the traditional doctor–patient interaction. It has a multitude of more features like 'Find My Inhaler' and 'Emergency Functionality' which we will be discussing exhaustively in further sections.

V. S. Shah (✉) · V. Musale
School of Computer Engineering and Technology, Dr. V. D. Karad MIT-WPU, Pune, India

V. Musale
e-mail: vinayak.musale@mitwpu.edu.in

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7_25

2 Literature Review

The survey is been carried out in references with different papers which were published by reputed journals and publishers in related domains. Throughout the survey, we discovered different approaches been used for building sort of similar technology and discovered their backlashes and how we can overcome those through proposed system implementation. Kikidis et al. proposed the idea of inculcating different smart IoT devices onto inhaler to increase the efficiency of the normal inhaler. The authors primarily focused on the creation of a sensing framework that can provide accurate information about health of patients and help their doctors to understand any possible difficulties that prevent patients from using their inhaled medication correctly using different sensors and devices [1].

Studying different sensors and their functionality was also a major part of the survey, did some findings on Arduino-based real-time air quality and pollution monitoring system. The development of an Arduino-based air pollution detector is carried by Roy et al. which combined a small-sized, minimum-cost sensor to an Arduino microcontroller unit. The reliable stability, rapid response recovery, and long-life features are some of the advantages of their proposed detector [2].

Kumar et al. developed android-based heart rate monitoring and automatic notification system where heart rate sensing is an objective of their developed project. The sensor and other required circuitry was unified into a small clip-like structure so that a patient suffering from heart disease wear it on the fingertip continuously and send notifications through short message service (SMS) over the cellular network using presented android application [3].

Smart asthma inhaler system was proposed by Deeksha et al. monitor the patient's health conditions, tracks the location, and the time inhalers are used so as to provide suitable medication to the patients once the reports are analyzed by the physicians involved [4].

Also, the research was carried on temperature measurement system using an LM35 temperature sensor, Arduino-based real-time wireless temperature measurement, and logger system. As presented in [5], wireless transmission of the temperature data occurs via a Bluetooth link.

Design of heart rate equipment based on Bluetooth communication on bike speedometers, from the voltage source, Arduino will supply the voltage for the pulse sensor and Bluetooth HC-05. The pulse sensor will be given voltage by Arduino of 5 V, and Bluetooth get voltage equal to 3.3 V. When all components are active it will be shown on LCD which is applied to handlebar bike as presented by Adam et al. [6].

Caleb Phillips et al. presented a reliable peripheral-named wrist-worn pulse oximeters where signal can be taken automatically from the wrist using sensors similar to those currently employed in existing wrist-worn devices. The WristO2 uses pulse oximeter and motion data to detect and reject unreliable data, which reduces the average error from 14.5 to 1.5% as mentioned [7].

3 System Overview and Objectives

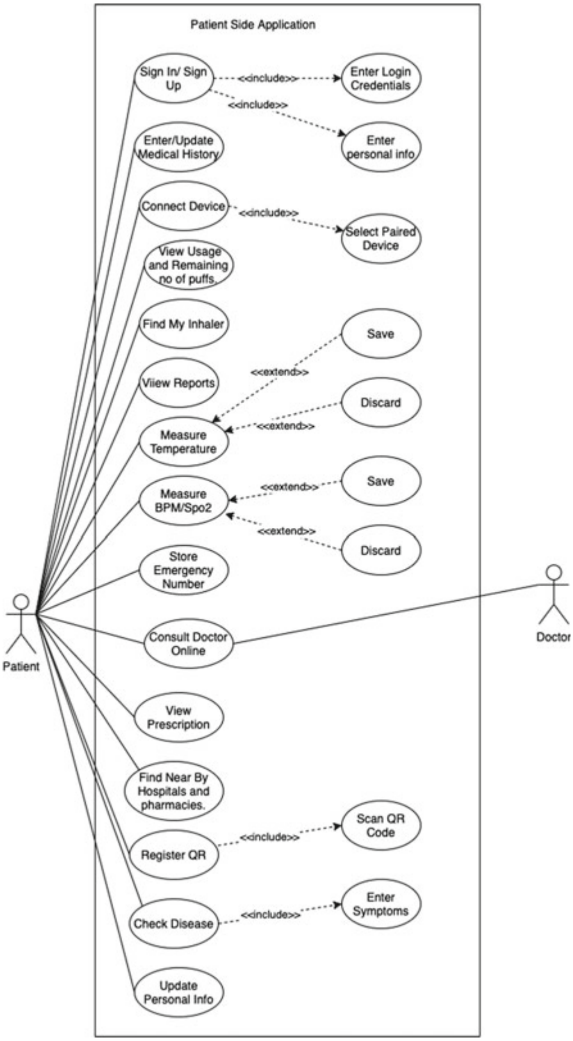
The proposed system consists of a retrofit hardware device developed on a Seeeduino Xiao microcontroller with a plurality of sensors and components encased together and two mobile applications, one for the patient side usage and the second for the doctor's side usage. The interaction between the patients and this hardware (hereinafter referred to as inhaler) takes place using the patient side application, which also allows the patients to do much more like using all the talked about functionalities and the doctors network wherein the patient can communicate with the doctors around the globe, with the ease of sitting at their homes; thus creating the need for the second application to facilitate the interaction between the doctors and patients.

The inhaler developed using the Seeeduino Xiao microcontroller encompasses an LM35 temperature sensor for body temperature measurement and a GY-MAX30100 oximetry sensor for pulse rate and Spo2 measurement. The patient is called to place his/her finger on the corresponding sensors during the measurement process, which can be initiated from the patient side application and requires the patient's finger to stay put until the measurement process is completed, and the obtained values are displayed on the patient side application. The inhaler also packs a buzzer, which is set ringing, whenever the patient is unable to find his/her inhaler in an indoor surrounding using the 'Find My Inhaler' functionality. The device has an air quality sensor that performs ceaseless monitoring of the surrounding of the patient and alerts them when the air quality has degraded, using the Groove Air Quality Sensor V1.3 installed in the inhaler. The inhaler houses two buttons, first for keeping the track of the inhaler usage and the other serves as an emergency button, which enables the patient to seek help in case of an emergency. The data transfer between the inhaler and the patient side application is carried out using Bluetooth and employs the HC-05 module for the same. Each of these inhalers has a unique QR code which makes the process of tracking down the patient easier in case the inhaler is lost somewhere. The inhaler is powered using a 1000mah Li-ion battery and can be charged using micro-USB chargers (Fig. 1).

The patient side application allows patient to interact with the inhaler where measured data are displayed and processed. The application requires the patient to sign in or sign up and is asked to enter all the necessary information at the time of signup like name, email, gender, and date of birth. The patient is also required to enter his/her medical history in detail on the application with the aim that the doctors can prescribe medicines after an exhaustive study of the patient's medical history when using the doctor's network functionality.

To use the inhaler-related functionalities, the patient first needs the patient side application to be connected to the inhaler using Bluetooth, which once completed the current usage count and remaining counts are displayed on the application and delivers a warning notification, whenever the remaining puffs are below 20. Each time the inhaler is used, the counter increments, and a report is pop-up with information about the air quality, humidity, temperature, current medication, etc., automatically recorded into the respective field, providing the patients with an option

Fig. 1 Patient side application use case diagram



to save or discard these reports. The patient side application enables the patients to measure their heart rate and Spo2 and are given a choice of saving or discarding these values. However, with saved values, it is viable to see a trend in these values and help predict various conditions by observing the changes in the heart rate and Spo2, which is also demonstrated using a graphical format in the generated reports. Another distinctive functionality of the patient side application is the 'Find My Inhaler' functionality which allows the inhaler to be found easily in case of an attack in an indoor environment, which is achieved by ringing the buzzer installed in the inhaler.

The application also demands storing an emergency number to be contacted in case of an emergency, i.e., when the emergency button on the inhaler is pressed. Under such circumstances, the saved number is apprised with the patient's current location, and using the location data, the nearest hospital is contacted. Another accommodating feature of this application is the 'Doctor's Network' which enables the patients, especially the elderly who find it onerous to visit the doctor physically every time. With this functionality, the patients can interact with several authorized doctors on the network through the means of text messages, PDF reports, images, etc., and experience consultation at the ease of sitting at home. The patients can contemplate the crucial data about the doctors like their work experience, education, specialization, and success rate which helps the patient in selecting an ideal doctor. The patient side application allows the patients to view the currently prescribed medication and also a history of all the previous prescriptions while enabling them to set alarms for the medicine and doctor's appointment.

The patient side application provides a feature, wherein the patients can locate nearby pharmacies and hospitals, by displaying ten records along with their contact information and directions, sorted on the basis of their distances. Furthermore, it allows the patients to input the symptoms they are experiencing and predicts the disease they may be experiencing which makes the process of prognosis easier. The application also requires the patient to register their contact information with the unique QR code present on the inhaler so that it becomes easy to track down the patient in case the inhaler is lost, which is done using the 'Register QR' functionality where the QR code is scanned, and the patient is prompted to enter their contact details, and the details are then registered for that particular QR code. This QR code when scanned by a third person takes him/her to a Webpage displaying the contact details of the owner/patient.

For the doctor's network functionality to operate, another application is required for the doctor's side which provides a platform for the doctors to interact with the patients available on the patient side application. This application requires the doctors to sign in with their provided login credentials; however, they don't have an option for signing up, as this can result in the formation of misleading accounts, and mishaps can occur. To counter this, the login details are generated by the admin after a meticulous inspection of the data provided by the doctors and are granted access only if all the documents produced are accurate. However, the doctors have the freedom to introduce themselves in terms of years of experience, specialization, patient testimonials, etc., in detail, which the patients can view that helps to create a sense of confidence in the doctor. The doctors on the other hand can examine the patient's information like age, gender, medical history, prescription history to get an idea about the patient and thus allows him/her to prescribe new medication effectively. The application allows the doctor to interact with the patients in the form of text messages, PDFs, images, and various other media types. The prescription can be updated by the concerned doctor which contains information about medicines and their dosage, tests to be performed if any, and the next appointment.

3.1 Objectives

The idea is to take the traditional handheld inhaler and transform it into a more advanced handheld device, ensuring that the usability and portability of the original inhaler are not compromised. Additionally, the proposed system has the following objectives:

1. Keep track of the usage and number of remaining puffs and forewarn the patient when the number of puffs remaining is fewer.
2. The sensor keeps a continuous track of the air quality and warns the patient when he/she enters an area with degraded air quality or in an environment with air quality where he/she has received an attack before. The sensor is not only restricted in its ability to sense the environmental gases but warns of strong scents as well like heavily applied perfume, incense stick which can plausibly trigger an attack.
3. The application allows locating the inhaler easily when required by using the 'Find My Inhaler' functionality, which sets the buzzer encased in the inhaler ringing.
4. The device allows the patients to monitor their heart rate, Spo2, and body temperature and warns the patient if the values lie outside the threshold range. Periodic analysis of these values can benefit to understand the trend in these values and predict and prevent a heart attack or similar condition.
5. Automatic/optional report generation feature generates a timely report with all the crucial data like air quality, humidity, usage per week, usage per month, current prescription and doctor, and a graph of the historic heart rate and Spo2 values which allows the doctor for a better prognosis and the patient for better precautions.
6. With the emergency button installed on the inhaler, a single press of the button activates the emergency functionality, wherein the patient's relatives are messaged with the patient's current location and the nearest hospital is contacted immediately.
7. The application has a 'Doctors Network' which revolutionizes the practice of doctor–patient interaction, by putting the patients at ease as they can now get a consultation from the doctors while sitting at their home and are not restricted in terms of where the doctor is situated, as any doctor around the globe can be consulted while providing the doctors with the ability to view and update their prescriptions online.
8. The application allows for a machine learning feature, wherein the patient can uncover the disease they might be having by admitting the symptoms they are encountering making the process of prognosis easier.
9. The application also has features like show nearby medical shops and hospitals and setting alarms for medications and doctor appointments.

4 System Design and Methodology

This section explains the working of different blocks in the system, the communication between them, and the technologies used.

A. *Hardware*

The hardware is developed using the Seeeduino Xiao microcontroller owing to its small size, which is a deciding factor for such handheld devices. The traditional choice of the microcontroller for such applications includes Arduino Nano or NodeMCU, but this microcontroller proves to be half the size, with all the required functionalities, making it suitable for this application (Fig. 2).

As already seen in the system overview, the hardware encases a plurality of sensors for tracking various values which include the air quality sensor, temperature sensor, oximetry sensor. Furthermore, the count value is also kept a track of and is stored in the EEPROM which is a more permanent storage solution so that the count value is not lost each time the device is restart. All these values are transmitted to the patient side application via Bluetooth, but owing to the numeric nature of all these values, an efficient key-value mechanism is used to map the numeric values with the corresponding fields.

Here ';' is used to split the data into keyword and their values; thus, the values can now be correctly understood by the application. For the heart rate, Spo2, and temperature measurement, the inhaler takes five readings and averages them out for better accuracy before sending it to the application. The following chart explains the communication between the inhaler and the application.

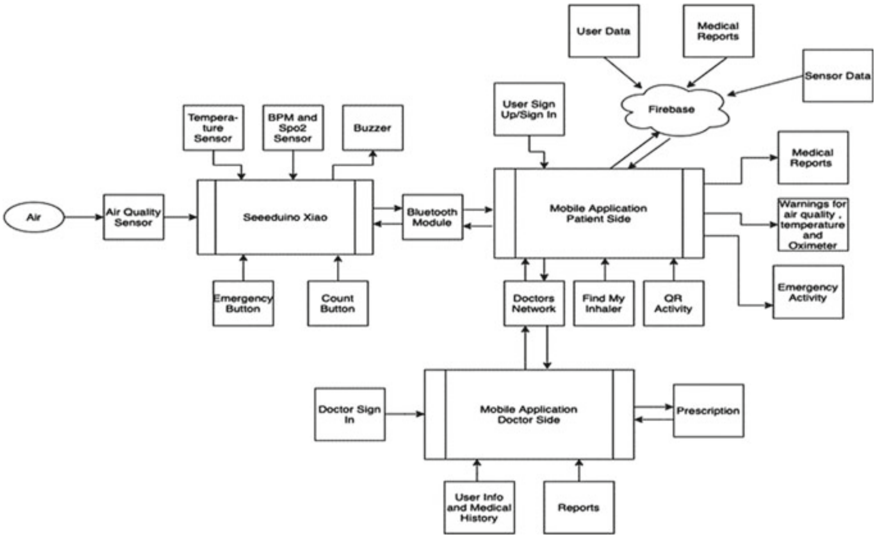


Fig. 2 System design overview

B. *Patient Side Application*

The application uses Firebase services for authentication of users, reading and writing values to the database, and storing the shared media. As seen in Table 1, the communication is carried out between the application and the inhaler using a key-value mechanism.

The application delivers warnings under circumstances like when the air quality has degraded or when the measured heart rate, Spo2, and temperature values are beyond the permissible limits. The doctor's network facilitates the patients to chat with the doctors and share media like PDF reports and images which is implemented using the Firebase real-time database and Firebase storage. The user signup data like name, email, and gender along with their medical history are also maintained using the Firebase real-time database for further processing. Figure 3 shows design of patient side application interface.

The noteworthy emergency functionality is employed using the Google Places API which enables the application to spot the nearest hospital and retrieve its contact information, and a call is placed immediately while notifying the emergency number with the current location of the patient. The same API is used to find the nearby hospitals and pharmacies, providing a list of hospitals and pharmacies along with their contact information and directions. The disease prediction functionality wherein the patients are required to enter their symptoms to predict the most plausible disease is implemented using the multinomial Naive Bayes algorithm from the sklearn package, and the Python code is integrated with the android environment using Chaquopy.

The model is trained using a dataset sourced from Kaggle which covers a wide range of 132 symptoms and 41 diseases with nearly 5000 records for better generalization of the model. The dataset also provides descriptions and a few precautions

Table 1 Communication between device and application

Message received from application	Action	Message sent from inhaler to application
Setup	Send count to inhaler	e.g., setup; 20;\n
temphigh	Start temperature sensor and send measured temperature value	e.g., temperature; 97.65;\n
BPM	Start oximeter sensor and send measured BPM and Spo2 value	e.g., oximeter; 87;98;\n
Reset	Reset count value and update EEPROM	e.g., initialize; 0;\n
aqi_update	Send air quality status to the application	e.g., aqi_update;Good;\n alert; The air quality has degraded, use a mask;\n
buzzerhigh	Starts ringing the buzzer	None
None	Activate emergency functionality in the application	e.g., emergency;\n
None	Update the count in the application	e.g., initialize;0;\n

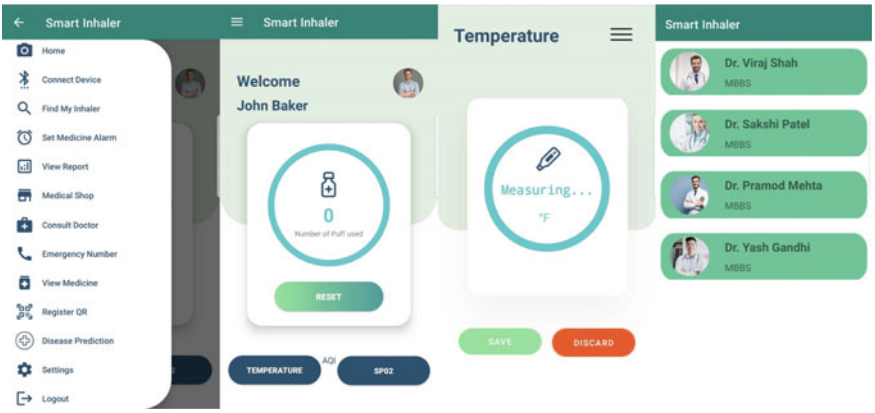


Fig. 3 Patient side application interface

for each of the diseases and is displayed to the patient which makes the process of prognosis and care easier. The doctor–patient prescription data are stored and updated using the Firebase real-time database while maintaining a history of all the prescriptions.

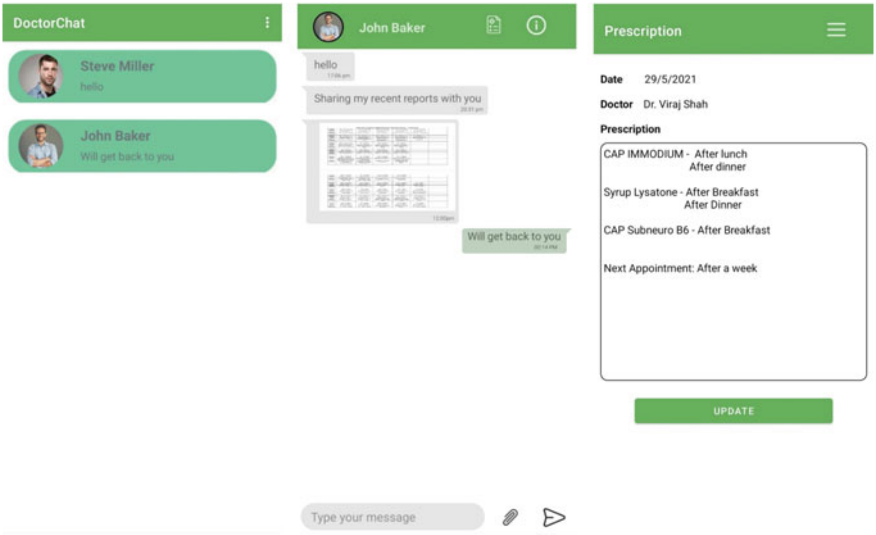
The BPM, Spo2, and temperature for each patient can also be stored in a date–time manner into the real-time database if the patient chooses to, and this historic data can be used to capture various trends in the value. The application also provides basic functionalities like forgot password, reset password, delete account, and update the profile using the Firebase authentication services.

C. *Doctor Side Application*

This application utilizes the same Firebase real-time database as the patient side application allowing for a centralized database between the two applications. This application enables the doctors to interact with the patients on the portal and view the patient information and medical history which is all retrieved from the Firebase real-time database, as this information is crucial for prescribing medicines. The doctor can examine the patient's prescription history and can accordingly update the prescription with the current date and time, which is then pushed to the database. All the media shared on the platform is stored on the Firebase storage and is also downloaded to the patient's/doctors mobile. The following Fig. 4 shows design of doctor side application interface.

5 Future Scope

Despite the fact that the current system is already a huge advancement over the existing systems, it can be tweaked in a few ways like:



hello

Sharing my recent reports with you

Will get back to you

Date 29/5/2021

Doctor Dr. Viraj Shah

Prescription

CAP IMMODIUM - After lunch
After dinner

Syrup Lysatone - After Breakfast
After Dinner

CAP Subneuro B6 - After Breakfast

Next Appointment: After a week

Fig. 4 Doctor side application interface

- Reducing the size of the inhaler with the development of smaller microcontrollers and sensors.
- Using GSM technologies that can allow the device to be traced even in outdoor conditions.
- Incorporating sensors like accelerometer and gyroscope can help predict a fall and emergency actions can be taken.

6 Conclusion

As asthma is a chronic and critical illness, an emergency can arise at any moment. Keeping a track of medication has also been an issue for the patient along with other factors and different changes that happen in the body. Keeping in mind all these issues, it was high time a device is developed to solve these problems. This system will help the patient to track the usage of the pump while monitoring the remaining number of puffs. Apart from this, the device will constantly monitor AQI and will alert the patient about certain AQI when they had an attack. In additional features like temperature and heart rate monitoring, patients can maintain a proper medical report. The medical report that is generated can be sent to different doctors for evaluation. In case of a severe attack, the patient can contact the nearest hospital as well as the emergency contact with just a push of a button.

References

1. Kikidis D, Konstantinos V, Tzovaras D, Usmani OS (2016) The digital asthma patient: the history and future of inhaler based health monitoring devices. J Aerosol Med Pulmonary Drug Delivery
2. Ahasan A, Roy S, Saim AHM (2018) Arduino-based real-time air quality and pollution monitoring system. Int J Innov Res Comput Commun Eng
3. Sadhukhan RK, Haque MM, Rahman N (2017) Android based heart rate monitoring and automatic notification system. IEEE
4. Anulaa M, Dayaghan Shanbhag D, Prasad MK, Roja AL (200) Smart asthma inhaler. IEEE
5. Chatterjee S, Chatterjee S, Gupta R (2017) Arduino based real-time wireless temperature measurement system with GSM based annunciation. IEEE
6. Faroqi A, Nuraeni E, Belawi H (2020) Design of heart rate equipment base on bike speedometer. IEEE
7. Phillips C, Liaqat D, Gabel M (2019) Reliable peripheral oxygen saturation readings from Wrist-worn pulse oximeters

Author Index

S. Bhalla et al. (eds.), *Proceeding of International Conference on Computational Science and Applications*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-0863-7