Lecture Notes in Electrical Engineering 724

Rayner Alfred
Hiroyuki Iida
Haviluddin Haviluddin
Patricia Anthony *Editors*

# Computational Science and Technology

7th ICCST 2020, Pattaya, Thailand, 29–30 August, 2020

Springer

# Lecture Notes in Electrical Engineering

## Volume 724

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

**China**

Jasmine Dou, Editor (jasmine.dou@springer.com)

**India, Japan, Rest of Asia**

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

**Southeast Asia, Australia, New Zealand**

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

**USA, Canada:**

Michael Luby, Senior Editor (michael.luby@springer.com)

**All other Countries:**

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

More information about this series at http://www.springer.com/series/7818

Rayner Alfred · Hiroyuki Iida ·
Haviluddin Haviluddin · Patricia Anthony
Editors

# Computational Science and Technology

7th ICCST 2020, Pattaya, Thailand,
29–30 August, 2020

Springer

*Editors*
Rayner Alfred
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu, Sabah, Malaysia

Haviluddin Haviluddin
Faculty of Computer Science
and Information Technology
Department of Informatics
Universitas Mulawarman
Samarinda, Kalimantan Timur, Indonesia

Hiroyuki Iida
Research Center for Entertainment Science
Japan Advanced Institute of Science
and Technology
Nomi, Ishikawa, Japan

Patricia Anthony
Department of Environmental Management,
Faculty of Environment, Society and Design
Lincoln University
Christchurch, New Zealand

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-33-4068-8 ISBN 978-981-33-4069-5 (eBook)
https://doi.org/10.1007/978-981-33-4069-5



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

# Preface

Computational science and technology is a rapidly growing multi- and interdisciplinary field that uses advanced computing and data analysis to understand and solve complex problems. The absolute size of many challenges in computational science and technology demands the use of supercomputing, parallel processing, sophisticated algorithms and advanced system software and architecture. The ICCST2020 conference provides a unique forum to exchange innovative research ideas, recent results, and share experiences among researchers and practitioners in the field of advanced computational science and technology. Building on the previous four conferences that include Regional Conference on Computational Science and Technology (RCSST 2007) and a series of International Conference on Computational Science and Technology (2014–2019), the Seventh International Conference on Computational Science and Technology (ICCST2020) program offers practitioners and researchers from academia and industry the possibility to share computational techniques and solutions in this area, to identify new issues, and to shape future directions for research, as well as to enable industrial users to apply leading-edge large-scale high-performance computational methods. This volume presents a theory and practice of ongoing research in computational science and technology. The focuses of this volume are on a broad range of methodological approaches and empirical references points including artificial intelligence, cloud computing, communication and data networks, computational intelligence, data mining and data warehousing, evolutionary computing, high-performance computing, information retrieval, knowledge discovery, knowledge management, machine learning, modeling and simulations, parallel and distributed computing, problem-solving environments, semantic technology, soft computing, system-on-chip design and engineering, text mining, visualization and web-based and service computing . The carefully selected contributions to this volume were initially accepted for oral presentation during the Seventh International Conference on Computational Science and Technology (ICCST20) held on 29–30 August 2020 virtually. The level of contributions corresponds to that of advanced scientific works, although several of them could be addressed also to non-expert readers. The volume brings together 47 chapters. In concluding, we would also like to express

our deep gratitude and appreciation to all the program committee members, panel reviewers, organizing committees, and volunteers for your efforts to make this conference a successful event. It is worth emphasizing that much theoretical and empirical work remains to be done. It is encouraging to find that more research on computational science and technology is still required. We sincerely hope the readers will find this book interesting, useful, and informative and it will give then a valuable inspiration for original and innovative research.

Kota Kinabalu, Malaysia — Rayner Alfred
Nomi, Japan — Hiroyuki Iida
Samarinda, Indonesia — Haviluddin Haviluddin
Christchurch, New Zealand — Patricia Anthony

# Keynote Speakers

Dr. Nilanjan Dey, Assistant Professor, Department of Information Technology, Techno India College of Technology, India

Dr. Suresh Manandhar, Director, Nepal Applied Mathematics and Informatics Institute, Nepal

Prof. Emeritus Dato' Dr. Tengku Mohd bin Tengku Sembok, Professor Emeritus of Computer Science, National Defence University of Malaysia, Malaysia

# Contents

# Building a Knowledge Graph of Vietnam Tourism from Text


**Phuc Do and Hung Le**

**Abstract** Most data in the world is in form of text. Therefore, we can say text stores large amount of the knowledge of human beings. Extracting useful knowledge from text, however, is not a simple task. In this paper, we present a complete pipeline to extract knowledge from paragraph. This pipeline combines state-of-the-art systems in order to yield optimal results. There are some other Knowledge Graphs such as Google Knowledge Graph, YAGO, or DBpedia. Most of the data in these Knowledge Graphs is in English. On the other hand, the results from our system is used to build a new Knowledge Graph in Vietnamese of Vietnam Tourism. We use the rich resources language like English to process a low resources language like Vietnamese. We utilize the NLP tools of English such as Google translate, Stanford parser, Co-referencing, ClausIE, MinIE. We develop Google Search to find the text describing the entities in the Internet. This text is in Vietnamese. Then, we translate the Vietnamese text into English text and use English NLP tools to extract triples. Finally, we translate the triples back into Vietnamese and build the knowledge graph of Vietnam tourism. We conduct experiment and discover the advantages and disadvantages of our method.

**Keywords** Knowledge graph · Google search · Triples extraction · Co-reference resolution · Natural language processing

## 1 Introduction

The information that we have nowadays is larger than it has ever been before. Most of the time, this enormous amount of data is text and come mostly in form of unstructured data. Text provides a quick and simple way to transform ideas from one person to

P. Do (✉) · H. Le
University of Information Technology,
Vietnam National University, Ho Chi Minh City, Vietnam
e-mail: phucdo@uit.edu.vn

H. Le
e-mail: hungle1abc@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_1

another. In this big dataset, there are knowledge hidden everywhere. We have yet to find the best way to transform this data into useful knowledge.

It is impossible for any person to read all the text in the world. This is the problem we want to address because we need to find the information which we want in this ocean of words without wasting time reading about unrelated subjects.

Text are written in natural language, which is complex due to its ambiguity. The same sentence can have two different meanings in two different paragraphs. Natural language is so flexible that we can use it in different contexts with ease, but this feature made it extremely hard for computer to understand.

The flexibility of natural language makes it impossible to define a set of rules that can cover all use cases of it. Instead, we have to use algorithms that can extract meaning of each sentence and collect the core information from it. The core information is called "facts", or "triples". Any fact can be expressed as a triple of form [Subject, predicate, Object], where Subject and Object are names for real world entities, and predicate is the relationship between these entities. An example triple is [Sesame, is_a, food]. These triples are the basic components of a Knowledge Graph [1].

In this paper, we present a system to build knowledge graph of Vietnamese tourism in Vietnamese. Unluckily, Vietnamese is a low resources language which only has a few NLP resources (software and large data set).

The method we used in this study is that we leverage the powerful resources of English language such as Google translate, Stanford parser, Co-referencing, ClausIE, MinIE. We use Google Search to find the text of specified entity, then we process and translate this bare Vietnamese Text to English. After that, we use English NLP tools to process English Text. Finally, we convert text back to Vietnamese to build the knowledge graph. The knowledge graph will contain a lot of facts about tourism in Vietnam.

In our study, we used NeuralCoref to solve co-reference resolution in the paragraphs [2]. Then we used MinIE [3] to extract triples from sentences. Finally, a graph database called Neo4j [4] is used to store the extracted triples.

We had chosen to solve this problem because it has tremendous applications. One of the applications is currently used by Google search engine. When a user searches for some keywords, Google Search display a box that shows summarized information from articles contains user's keywords.

In this study, we have following contributions:

- We develop a system to utilize the NLP tools of rich resource language like English to extract the triples from text in low resource language like Vietnamese Text. Our research can be a typical application for all low resource languages.
- We design a pipeline containing sequential steps of NLP tools to build a knowledge graph of Vietnamese Tourism in Vietnamese.
- We conduct the experiment and discover the advantages and disadvantages of our proposed method. This result will be a guideline for users who want to build a non-English knowledge graph.

The rest of paper is organized as follows: Section 2 presents the Related Work of our research. Section 3 presents our methodology to solve the problem. Section 4

presents the implementation of our proposed system. Section 5 presents the experiment and the advantages and dis-advantages of our proposed methods. Finally, we conclude what we did and discuss about the future work.

## 2 Related Work

Google Knowledge Graph, YAGO, and DBpedia are the most well-known knowledge graphs.

Google Knowledge Graph was built in 2012 by Google. It provides direct information quickly by using the relationships between words and concepts from the query. It makes use of user behaviors, related entities and relationships.

YAGO is an open source knowledge base which was developed at the Max Planck Institute for Computer Science. This knowledge base contains more than 10 million entities and more than 120 million facts about these entities. The information was extracted from Wikipedia, WordNet, and GeoNames [5]. This knowledge base is the favorite data source of researchers who interested in testing new ideas on graph database.

DBpedia is a project which was created to extract structured content from Wikipedia articles [6]. The English version of DBpedia knowledge base describes 4.58 million things, ranging from many different topics like persons, places, species, disease, etc. DBpedia extracts information based on Wikipedia article structures and link them together with an extraction manage. Similar to YAGO, DBpedia structured contents can be query using SPARQL and it is free to use.

## 3 Methodology

### *3.1 Definition*

There is no formal definition of knowledge graph. We consider knowledge graph as a graph where nodes are real world entities and edges are relationships between them. Moreover, knowledge graph also contain rules to enable reasoning to infer new knowledge from existent triples of knowledge graph.

### *3.2 Extract Triples of Knowledge Graph from Text*

To conduct the research, we first looked for other systems that are trying to solve somewhat similar problem. Out of the systems that we saw, a few of them was really stood out. Those are ClausIE, MinIE, NeuralCoref in English, and Neo4j graph database.

We proceeded to combine these systems to build an end-to-end pipeline that can take a paragraph as input and return the knowledge graph as output. This knowledge graph is store in a graph database for later processing. Finally, we ran some experiences, discussed the results of our pipeline and presented some ideas we have moving forward.

### *3.3 Introduction to ClausIE and MinIE*

ClausIE is one of the systems that were built to solve the task of Open Information Extraction (OIE) [7] in English. It consists of two separated steps. The first step is to detect "useful" information from the given sentence. This means ClausIE decides what information is expressed in the sentence, how to identify it, and how much of it worth keeping. The second step is to identify the sentence's representation. This is the part where ClausIE decides what is the form of the relation, should it use triples or n-ary proposition to generate representation of the information in text.

In order to detect useful information, ClausIE makes use of dependency parsing. It uses dependency parsing to detect the set of "clauses" of each sentence. No training data is needed for ClausIE to work properly. After the sets of clauses have been found, ClausIE's second step is to generate propositions for each clause based on the type of the clause. In the Table 1, the first clause pattern, "Tom" is the subject and "laughed" is the intransitive verb. Similarly, in the 7th clause pattern, "Tom" is the subject, "put" is the complex-transitive verb, "his computer" is the direct object, and "down" is the complement.

Though ClausIE achieves high precision and recall, it tends to produces overly-specific extractions. Therefore, MinIE was built on top of ClausIE to generate more useful and semantically richer extractions. For its extractions to be more compact, MinIE uses annotations for capturing the context of an extraction. These annotations represent information about polarity, modality, attribution, and quantities. MinIE also identifies and removes parts that are considered over-specific.

**Table 1** Basic clause patterns and their examples

| Clause patterns | Example sentences |
|---|---|
| $SV_i$ | Tom laughed |
| $SV_eA$ | Tom studied information systems |
| $SV_cC$ | Tom is a student |
| $SV_{mt}O$ | Tom likes books |
| $SV_{dt}OA$ | Alice gave tom a cup of coffee |
| $SV_{ct}OA$ | Alice taught Tom system design |
| $SV_{ct}OC$ | Tom put his computer down |

*S: Subject, V: Verb, C: Complement, O: Direct object, A: Adverbial, $O_i$: Indirect object, $V_i$: Intransitive verb, $V_c$: Copular verb, $V_c$: Extended-copular verb, $V_{mt}$: Monotransitive verb, $V_{dt}$: Ditransitive verb, $V_{ct}$: Complex-transitive verb

### 3.4 Co-reference Resolution and NeuralCoref

Co-reference resolution meaning finding all the words that refer to the same entity in a given piece of text. Let take a look at the following paragraph from Wikipedia: "VNUHCM-University of Information Technology is a public university located in Ho Chi Minh City, Vietnam. Although its name is about information technology, this university teaches many computer studies."

When human read the above paragraph, we can easily know that "its" and "this university" are referring to "VNUHCM-University of Information Technology". An effective co-reference resolution system should be able to do the same thing. Figure 1 shows that NeuralCoref can correctly determine the entities and their antecedences in our example paragraph.

### 3.5 Store Knowledge Graph in Neo4j Graph Database

Neo4j is a Graph Database management system. It was built to efficiently store, handle, and query highly-connected data. It has a powerful and flexible data model, so it is good choice to store semantic triples. A node in the graph could be a subject or an object in the triple, and the relationship between two nodes is the predicate between the subject and the object.

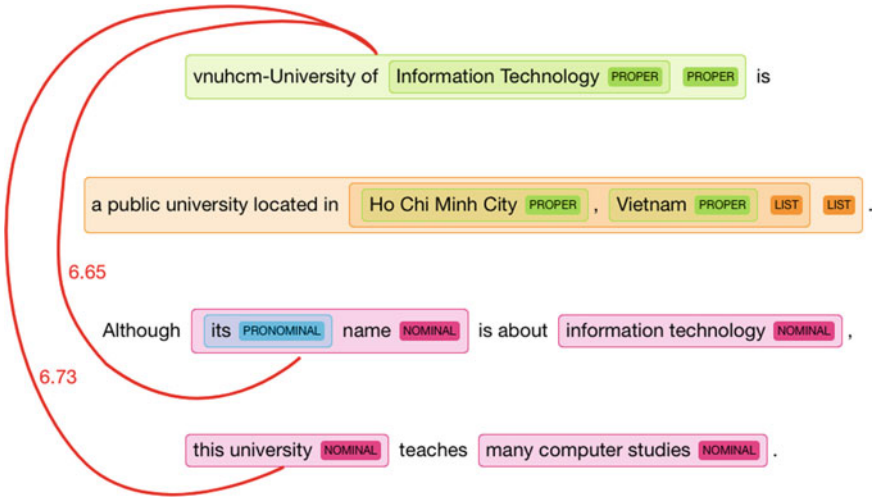



**Fig. 1** Results from NeuralCoref with our example

### 3.6 The Pipeline of Proposed System

The ultimate goal of our research is to build a system which can receive a paragraph written in natural language and create a Vietnamese Knowledge Graph. Figure 2 shows the architecture of our pipeline.

The first step in this pipeline is the collection of English paragraphs related to our topic. We tried to find paragraphs related to tourism in Vietnam because that is our focus. We choose English Text because of two main reasons. The first reason is the amount of text written in English is much larger than the amount of text written in Vietnamese. The second reason is processing text written in Vietnamese has not become as good as processing text written in English. By using English text as input, we can take advantage of the tools that have already been developed for years to process text. As we can see later on in this pipeline, with Google Translate, all the text will be translated to Vietnamese. This method allows us to take advantage of existing tools while being able to have the results in Vietnamese for later use.

The next component is the co-reference resolution followed by triple extraction component. Combine the two components give us more accurate results than using each of them individually. After the triples are extracted, we use Google Translate API to translate them into Vietnamese. Since the triples contain only phrases, not whole sentences, translation systems will do a generally good job. In cases where the phrases of Vietnamese Text are not translated correctly, an expert can step in and edit the translations directly.

The next component is an entity mapping suggestion component. This component uses Jaccard's similarity algorithm and a dictionary in order to give suggestions about the types of entities and relationships.

Finally, we stored the result in Neo4j.

### 3.7 The Structure of the Knowledge Graph

In this research, we tried to collect data that fall into one of the schemas in Table 2. We intended to build a knowledge graph of Vietnam tourism [8]. In Table 2, we use English and Vietnamese to describe the head, the predicate and the tail of triples of our knowledge graph.

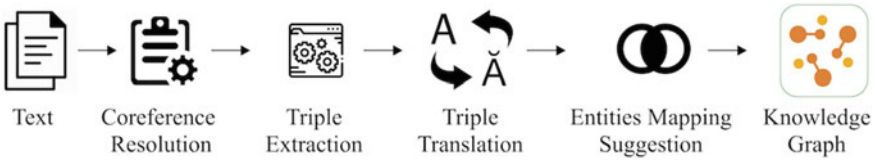



**Fig. 2** Knowledge Graph construction pipeline

**Table 2** Types of entities and relationships

| Head type | Predictate type | Tail type |
|---|---|---|
| Landscape (Thắng cảnh) | IN (TOẠ LẠC TẠI) | Place (Địa danh) |
| Festival (Lễ hội) | FESTIVAL IN (LỄ HỘI TẠI) | Place (Địa danh) |
| Hero (Anh hùng dân tộc) | WAS BORN IN (SINH RA TẠI) | Place (Địa danh) |
| Dish (Món ăn) | SPECIAL DISH AT (MÓN ĂN ĐẶC SẢN CỦA) | Place (Địa danh) |
| Folk song (Dân ca) | TRADITIONAL SONG OF (DÂN CA CỦA) | Ethnic Group (Dân tộc) |
| Musical instrument (Nhạc cụ) | INSTRUMENT OF (NHẠC CỤ CỦA) | Ethnic Group (Dân tộc) |
| Ethnic Group (Dân tộc) | LIVING IN (SINH SỐNG TẠI) | Place (Địa danh) |

Figure 3 clearly shows the structure of our Knowledge Graph. The nodes represent the types of entities in our Knowledge Graph, and the arrows represent the types of the relationships between them.

## 4 System Implementation

The implementation of the experimental system is as follow: (i) On the Client: just install a browser like Chrome or Firefox; (ii) On the Server: install Docker, set up environment variables and run the command "docker-compose up –build"; (iii) Go to http://127.0.0.1:5000 to test the system.

In order to generate the knowledge graph from paragraphs, our system has the following services as described in Table 3.

Detail descriptions and algorithms of the services of our system are described below.

**Table 3** The systems' services

| No. | Services | Description |
|---|---|---|
| 1 | Co-reference resolution | Resolved mentions of same entities in the paragraph |
| 2 | Triples extraction and translation | Extract triples from the paragraph |
| 3 | Types recommendation | Translate the triples into Vietnamese and recommend the type of entities/relationships |

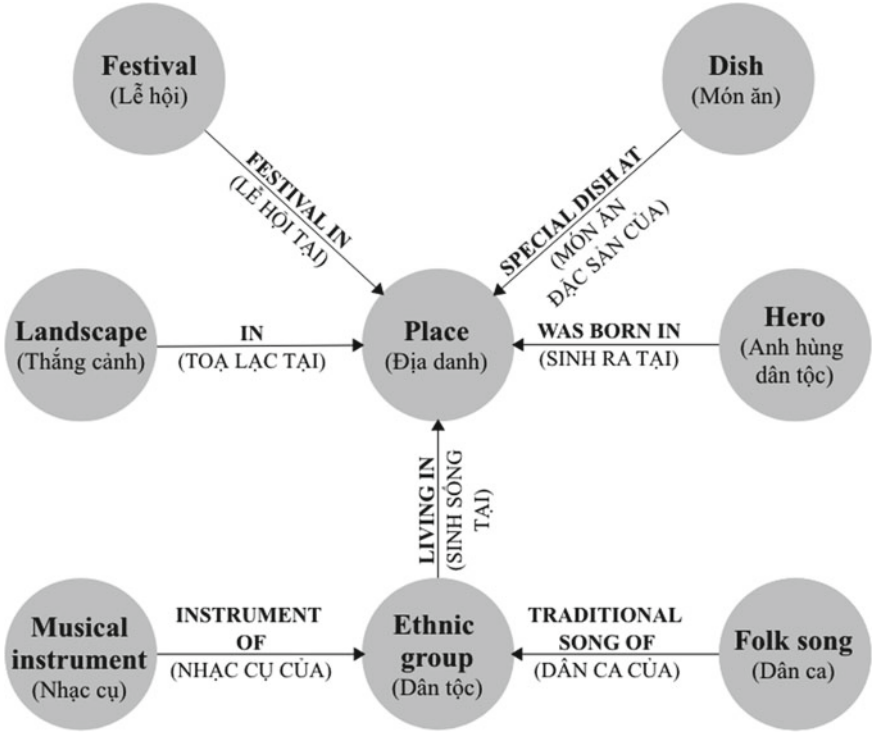



**Fig. 3** Structure of the Knowledge Graph

*(a) The "Co-reference Resolution" service:*

The *Co-reference Resolution* service is the service that take the responsibility to resolve entity co-occurrences. Within this service, we used NeuralCoref library to extract mentions and their duplications.

The input of this service is a normal paragraph, and the output of this service the resolved paragraph. The service did this by replacing the mentions of the same entity with the first occurrence of that entity in the paragraph. The details algorithm of the *Co-reference Resolution* service is described in Algorithm 1.

**Algorithm 1** The Co-reference Resolution algorithm

**Input:** Normal paragraph.
**Output:** Resolved paragraph.

1: Run NeuralCoref with the input paragraph to find mentions of the same entities in the paragraph.
2: Replace all the mentions of the same entities with the first occurences of those entities in all sentences of the paragraph.
3: Return the resolved paragraph.

*(b) The Triples Extraction and Translation service*

After we have the resolved paragraph from the *Co-reference Resolution* service, the next step is to extract useful triples to create our Knowledge Graph. This service takes each sentence of the resolved paragraph as its input and produces triples that the sentence contains.

The service detects triples by analyzing the English clause types. This analysis was done within ClausIE. MinIE then eliminates triples that are consider too specific and adds annotation to the triples. Our service then go a bit further and translate these triples into Vietnamese with human's verification. The result is a list of triples in English and Vietnamese along with their polarity and modality. Algorithm 2 describes how *Triples Extraction and Translation* service works.

**Algorithm 2** The Co-reference Resolution algorithm

**Input:** A sentence from the resolved paragraph.
**Output:** All the triples (in English and Vietnamese) that the sentence contains.

1: The sentence is processed with MinIE to produces triples that that sentence contains along with their polarity and modality.
2: An expert can change the translations at this step directly from the browser.
3: Return the list of triples that the sentence contains.

*(c) The Types Recommendation service*

The *Types Recommendation* service will recommend the type that an entity or a relationship is supposed to have. The list of possible type in this research is as described in Table 2.

We recommend the type by using VNCoreNLP [8] to segmentate Vietnamese words from the triples first, then compare these words with our dictionary using the Jaccard similarity algorithm to determine the type of the entity or relationship. Algorithm 3 is the explanation of this service.

**Algorithm 3** The Types Recommendation algorithm

**Input:** A Vietnamese triple.
**Output:** Recommended types of entities and the relationship of the triple.

1: The phrases of the triple is seperated into words with VNCoreNLP Word Segmentation [9, 10].
2: The words are compared with our dictionary using Jaccard similarity algorithm with the coefficient is 80
3: If their is no recommended type found, the returned type will be "UNKNOWN".
4: Return the recommended types for the triples.

## 5 Experiment and Discussion

In our system, the paragraphs are obtained through two sources: Google Search and Wikipedia. First, we collected the list of entities (written in Vietnamese) that we are interested in. Next, we either run these entities with our questions using Google Search to get the desired paragraphs about the entities; or we get the summary paragraphs of our entities by Python's Wikipedia API. Finally, these paragraphs are translated into English and feed to our system.

In this section, we will show the results of our system when we process the following paragraph which was taken from Wikipedia:

"$Dà$ Lạt city is the capital of Lâm $Dồng$ Province in Vietnam. The city is located 1,500 m above sea level on the Langbian Plateau in the southern parts of the Central Highlands region. Da Lat is the most popular tourist destination in Vietnam."

For clarity, we use a Text in English. Normally, Text is in Vietnamese and is translated to English by Google Translator and verified by man.

### *5.1 Results from the Co-Reference Resolution*

For this paragraph, the system understood the paragraph correctly. The mention "The city" in the second sentence of the paragraph is replaced with the entity "Da Lat". Fig 4 shows the resolved paragraph in our system.

### *5.2 Results from Triples Extraction and Translation*

For our example paragraph, the system extracted and translated a total of seven triples. They are listed in Table 4.

As the results shows that, some of the triples are wrong (No. 2), some of them are useless (No. 5), but some are pretty useful (No. 1, No.4, No.6) for our Knowledge Graph.

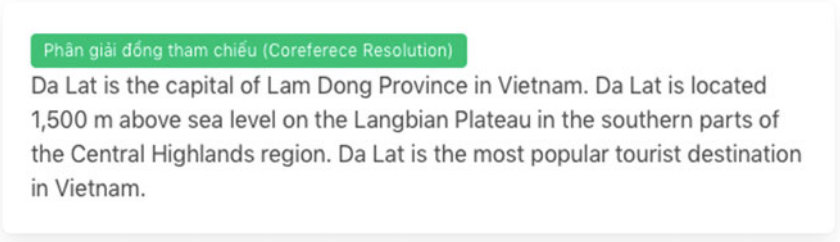



**Fig. 4** Resolved paragraph

**Table 4** Extracted triples from the example paragraph

| No. | Subject | Predicate | Object |
|---|---|---|---|
| 1 | Da Lat | Is capital of | Lam Dong Province |
| 2 | Da Lat | Is capital in | Vietnam |
| 3 | Da Lat | Is | Capital |
| 4 | Da Lat | Is located | 1,500 m above sea level on Langbian Plateau in southern parts of Central Highlands region |
| 5 | Da Lat | Is | Located |
| 6 | Da Lat | Is most popular tourist destination in | Vietnam |
| 7 | Da Lat | Is | Most popular tourist destination |

The triple No. 7 in Table 4 is translated into the Vietnamese triple: (“Đà Lạt”, “là hầu hết các du lịch nổi tiếng đích trong”, “Vietnam”). This is not the most accurate translation but as Google Translate get smarter, we can expect to get a better translation.

## 5.3 *Results from the “Type Recommendation” Service*

Fig 5 shows the recommended types for the triple No. 7 in Table 4.

The system can identify the entity ““Đà Lạt”” as a Place (“(“Địa danh”)”). For the other entity and the relationship, the system could not guess the type. However, the expert can step in and set the types of the entity/relationship directly, therefore, the system will be able to provide more suggestions over time.

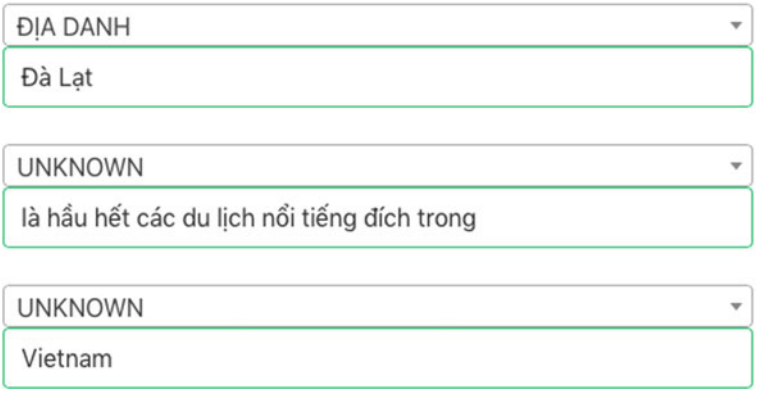



**Fig. 5** A triple with its recommended type

### *5.4 Discussions*

Through our experiments, we had found some advantages as well as some dis-advantages in our proposal system.
**Advantages** Our approach presents the following advantages:

- We can stay away from the complexity of the sub-problem which other tools have been trying to solve. Instead, we focused on the demands of our specific system such as the entities' types or the structure of our knowledge graph.
- By utilizing existing NLP tools, we were able to build our system quickly.

**Dis-advantages** Our approach, however, still has some cons:

- Although we had design our system so that we can easily replace any component with a better version, our system still depends on other NLP tools.
- Translations from Google Translate are not always accurate, and we lost some of the native features and characteristics that only exist in Vietnamese.

## 6 Conclusion and Future Work

In this paper, we proposed a system that can build a knowledge graph in Vietnamese of Vietnam Tourism. Our system is assembled by taking advantages of state of the art components of English like co-reference resolution, open information extraction, word segmentation, etc. Each of these components can also be further optimized independently. We hope that this system could be the baseline system that future systems in this domain can compare to.

In the future, we would like to use Deep Learning to exploit information from text.

**Acknowledgements** This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCMC) under the grant number DS2020-26-01.

## References

1. Ehrlinger L, Woss W (2016) Towards a definition of knowledge graphs
2. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) Transformers: state-of-the-art natural language processing. ArXiv
3. Gashteovski K, Gemulla R, Del Corro L (2017) MinIE: minimizing facts in open information extraction. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2630–2640
4. Webber J (2012) A programmatic introduction to Neo4j. In: Proceedings of the 3rd annual conference on systems, programming, and applications: software for humanity, pp 217–218
5. Suchanek F, Kasneci G, Weikum G (2007) YAGO: a core of semantic knowledge. In: 16th international world wide web conference, WWW2007, pp 697–706

6. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes P, Hellmann S, Morsey M, Van Kleef P, Auer S, Bizer C (2014) DBpedia—a large-scale. Multilingual knowledge base extracted from Wikipedia, semantic web journal
7. Corro L, Gemulla R (2013) ClausIE: clause-based open information extraction. In: WWW 2013—proceedings of the 22nd international conference on world wide web, pp 355–366
8. Do P (2019) SparkHINlog: extension of sparkDatalog for heterogeneous information network. J Intell Fuzzy Syst 37(6):7555–7566
9. Vu T, Nguyen DQ, Nguyen D, Dras M, Johnson M (2018) VnCoreNLP: a vietnamese natural language processing toolkit. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations, pp 56–60
10. Vu T, Nguyen DQ, Nguyen D, Dras M, Johnson M (2017) From word segmentation to POS tagging for Vietnamese. In: Proceedings of the 15th annual workshop of the Australasian Language technology association, pp 108–113

# Technology Adoption Models: Users' Online Social Media Behavior Towards Visual Information


**Irma Syarlina Binti Che Ilias, Suzaimah Ramli, Muslihah Wook, and Nor Asiakin Hasbullah**

**Abstract** Technology Adoption Model is used in various technology fields to understand and predict users' intentions and behaviors. However, the Technology Adoption Model used in Social Media, which explains users' intentions and behaviors needs to be investigated. Nevertheless, there is little understanding of users' intentions and behaviors towards visual information, which plays an important role in effective communication. This study reviewed a considerable amount of past studies on the use of the technology adoption model by users' online social media behaviors towards visual information. Based on the literature survey from FOUR (4) databases; ACM, IEEE, Scopus, and Science Direct; TWELVE (12) articles have been reviewed. The study found that Uses and Gratifications Theory (UGT) is the most adopted model due to the motivation mechanism applied. Most importantly, the review managed to discuss the models, factors, visual information, and methods in relation to users' social media intention and behavior. An Integrated Adoption Model could be developed to examine the consequences of the technology adopted to create a holistic understanding of how technology influences the users' intentions and behaviors towards visual information. This is one of the recommendations presented at the end of this research for the reference of future scholars.

**Keywords** Technology adoption models · Behaviors · Visual information · Social media

I. S. B. C. Ilias (✉)
Universiti Kuala Lumpur, 50250 Kuala Lumpur, Malaysia
e-mail: irmasyarlina@unikl.edu.my

S. Ramli · M. Wook · N. A. Hasbullah
Universiti Pertahanan Nasional Malaysia, 57000 Kuala Lumpur, Malaysia
e-mail: suzaimah@upnm.edu.my

M. Wook
e-mail: muslihah@upnm.edu.my

N. A. Hasbullah
e-mail: asiakin@upnm.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_2

**Table 1** EIGHT (8) models

| Model | Year |
|---|---|
| Social Influence Theory (SIT)_ | 1958 |
| Uses & Gratification Theory (UGT) | 1973 |
| Theory of Reasoned Action (TRA) | 1975 |
| Social Identity Theory (SIDT) | 1979 |
| Theory of Planned Behavior (TPB) | 1985 |
| Technology Acceptance Model (TAM) | 1989 |
| Social Cognitive Theory (SCT) | 1999 |
| Unified Theory of Acceptance and use of technology 2 (UTAUT2) | 2012 |

# 1 Introduction

A variety of models and frameworks were introduced to describe the adoption of technologies' models by users [1]. Several studies have utilized these conventional frameworks and combined previous models or added new constructs to existing models to carry out their researches [2]. In social media studies, there are several models that have been adopted to identify the factors of users' intentions and behaviors towards the use of technology adoption model. Table 1 shows the EIGHT (8) models discussed in this paper.

This paper aims to identify the models used in users' behaviors towards visual information in social media. Each model discusses the types of visual information, behaviors, factors, social media types and methods used. The paper structure are as follows: Sect. 2 describes the concepts of visual information, social media, users' online social media behaviors and the TWELVE (12) reviewed articles; Sect. 3 discusses on the Result and Findings; Sect. 4 concludes the paper; and Sect. 5 provide recommendations for future works.

# 2 Materials and Methods

## *2.1 Visual Information*

Visual information includes still photos, films, videos, sound recordings, graphic designs, visual aids, model displays, display systems and services, and the support processes. Recent statistics shows that social media posting with images has more engagement than those without [3].

## 2.2 Social Media

Social media is a combination of networking sites which allow users to communicate with one another easily, efficiently, and effectively [4]. It is about transforming (one-to-many) broadcasts into (many-to-many) dialogs which vastly contrast with more mainstream media sources such as televisions and books that serve information to mass audiences but do not allow user behavioral reactions onto their content [5]. Social media, on the other hand, requires users to create an online profile which enable users to connect with each other by notifying others of their latest content. What started out for private needs migrated into industry. Industries use social media to promote goods, create brand images and increase their main website traffic. Facebook, Instagram, Pinterest, Snapchat, and Twitter are the preferred visual information social media [6–10].

## 2.3 Social Media Behavior

Social media have facilitated the users with various public behaviors as a social gesture in supporting each other [11]. Users behave through Comment, Like, Post, Reactions, Share, and Tags [12–17]. This action does gratify users' social needs to maintain relationships [18].

## 2.4 Technology Adoption Models

Modeled behavior is often imitated more frequently if it is socially recompensed. Comment, Like, Post, Reactions, Share, and Tags are shown to all designated users in social media, which serve as a social reward and promote the recognition and adoption of model behaviors and attitudes In this paper, EIGHT (8) models are given to show the concepts used in adoption process. The models presented here are shown in Appendix. Technology Adoption Models used in Users' Online Social Media Behaviors.

### 2.4.1 Social Cognitive Theory (SCT)

Social cognitive theory (SCT) explores the user potentials of observer learning via behavior modelling, which also plays a key role in motivation and self-effectiveness [19–21]. Dhir, Kaur and Rajala [22] selected SCT to consider the Facebook photo-tag's possible behavior. TWO (2) significant measures were established by SCT: self-efficacy and expected outcomes. This study included 768 adolescent Facebook users. For data processing, SPSS 21.0 and AMOS 21.0 were chosen. Confirmatory

Factor Analysis (CFA) in AMOS was used to test the validity and reliability of the measurement model. The method of estimating maximum likelihood (ML) is used because of the robustness of the large data samples. However, results showed that direct influences of expected outcomes and self-efficacy on user intentions was not the main factors for the users to photo tagging in Facebook.

Peng et al. [23] used SCT to examine impacts of Instagram images on the intention of use and behavioral reactions. The SCT stresses in the ability of the human population to be observed through behavioral modelling in which individual's motivation (self-improvement, self enhancement, intention), model attractiveness, affective response (pleasant affection) and self-efficacy are important roles used in [20] research. A total of 1587 people was recruited as volunteers. The ANOVA, Scheffee, CFA and EFA were used for estimation of shared variability on all items in the measuring model. However, results showed that model attractiveness would influence a user's intention to self-improve and response affectively to the images shared on Instagram.

### 2.4.2 Social Identity Theory (SIDT)

Tajfel and Turner [24] proposed Social Identity Theory (SIDT) which elaborated on the users' self-conception of their relationship to another person or group [25, 26]. SIDT used by Nedra et al. [27] described how user intentions can be shaped by cognitive, affective, and evaluative ways in which each component can have significant effects on the intention in relations to posting photos on Instagram. SPSS and the CFA using SEM through AMOS have reported that the TWO (2) methods used in this study were quantitative and qualitative ones that involved 359 respondents. This finding supports the positive impact of the intention of SIDT (cognitive, affective and evaluative) to use Instagram for the posting of images.

### 2.4.3 Social Influence Theory (SIT)

Social Influence Theory (SIT) is defined as to what extent a person expects from the persons who are important to its behavior [28, 29]. Oliveira et al. [16] focused on users' intention to share contents on Facebook via SIT. Contents shared through Facebook does give impacts on users' behavior which then convinces and influences larger audiences in a more beneficial, effective and powerful manner. There were THREE (3) constructs involved; identification, internalization and compliance. The data was collected from a group of 687 users, aged from 20 to 55 years old. The study employed PLS SEM path modelling, to estimate and test the linkage between constructs. Results showed that internalization was the most important way to explain how users take a view of others as evidence of the reality and as part of the values and beliefs of each user when they share contents on Facebook.

### 2.4.4 Technology Acceptance Model (TAM)

Technology Acceptance Model (TAM) explore the factors affecting the individuals behaviors in using technology or computer systems [30–32]. Allam et al. [28] used TAM to postulate that the behavioral intentions of the users to tag their images are dictated by their attitudes to use Facebook. The online survey yielded 187 responses where PLS SEM method is used to validate latent variable for data analysis. Results showed that the perceived usefulness did influence users' attitude on social tagging tools.

Nedra et al. [27] integrates TAM model to explain users' behavior and intention towards posting photos in Instagram. An online survey was conducted on 338 participants where SPSS, CFA, SEM, AMOS are used for quantitative, while for qualitative, individual interviews were applied. The results indicated that photo-posting in Instagram was positively influenced by users' perceived ease of use.

Shao and Kwon [30] analyze the TAM model based on Facebook Reactions, a good example of complex social feedback systems. The construct is perceived value and ease of use. The survey was completed by a total of 432 participants, age 18 and/or above. The structural modeling was performed by using PLS-SEM and PLS-MGA. Results showed that perceived usefulness was the primary predictors of the intention of users to use Reactions, which further influenced the real Facebook usage.

### 2.4.5 Theory of Planned Behavior (TPB)

Theory of Planned behavior (TPB) states that users' intentions are the immediate determinants of their behavior [33, 34]. Lowe-Calverley and Grieve [14] aimed to investigate the factors affecting the intentions of users to post photos on the Facebook platform using TPB. The sample consisted of 151 Australian respondents. To allow shared variances among predictors, a Hierarchical Multiple Regression Analysis (HMRA) was used. Results show that user's intent to post images on Facebook are due to attitudes, subjective norms and narcissism factors.

Kim et al. [35] used TPB to forecast factors affecting Instagram photo posting. There was a total of 89 respondents from Instagram, and a path analysis was performed using AMOS 22 on the hypothesized model. Results showed that attitude, subjective norm, perceived behavioral control, and narcissism were the main factors of users' intent to post photos on Instagram.

### 2.4.6 Theory of Reasoned Action (TRA)

Theory of Reasoned Action (TRA) is a model for behavioral intention prediction [28, 36]. By using TRA, Lee et al. [37] focused on the Facebook "like", to see how social and technological factors influence users' attitudes and behaviors; social norms, social presence, perceived ease, and perceived usefulness. Online survey data were examined using Exploratory Factor Analysis (EFA), Canonical Correlation Analysis

(CCA), and Hierarchical Regression Analysis (HRA) for 213 respondents. Results showed that users' attitude towards "like" was positively influenced by subjective norms and ease of use.

### 2.4.7 Uses and Gratifications Theory (UGT)

The Users and Gratifications Theory (UGT) explains how people use media and explores various gratifications which drive their use [33, 38]. Lowe-Calverley and Grieve [11] use UGT to examine how Facebook users participate in image posting and liking behaviors. Using thematic analysis, 203 respondents were evaluated prior to posting and 195 respondents prior to liking photos on Facebook. Results showed that the 'audience' aspect is the key factor for image sharing on Facebook, while friends and enjoyment of content were the key factors affecting 'liking' actions.

Shao and Kwon [30] examine mechanisms of gratification that underlie the use of UGT in Facebook Reactions. In total, there were 432 participants who completed the survey where PLS-SEM and PLS-MGA have been selected data analysis. Results showed that enjoyment and expression were the main factors of user intention to use Facebook Reactions. While Malik et al. [39] applied UGT to examine the gratifications of users when sharing photos on Facebook. SPSS 21.0 and EFA with MLE algorithm were used to examine 368 respondents answer. Results showed that the level of disclosure and social influence of Facebook's photo sharing gratifications increased with an increase in users' age.

In addition, motives in UGT were studied by Lee et al. [37]. Exploratory Factor Analysis (EFA), Hierarchical Regression Analysis (HRA) and Canonical Correlation Analysis (CCA) were used to examined 213 respondents from whom data was collected. Results showed that when users find the content enjoyable, they would click "like" on the Facebook, which immediately then spreads out. Bij de Vaate et al. [6] applied UGT to study gratifications and behaviors of users in posting photos via Facebook. There were 224 respondents involved in the surveys and the data were analyzed using Multiple Linear Regression Analysis (MLRA). Results showed that main gratifications factors in Facebook photo posting were entertainment and retention of moment.

### 2.4.8 Unified Theory of Acceptance and Use of Technology Version 2 (UTAUT2)

Unified Theory of Acceptance and Use of Technology Version 2 (UTAUT2) involve the adoption of the intention to "continue to use the technology" i.e. to engage users who have used the technology [37]. The study by Dhir et al. [22] used many UTAUT2 factors related to photo tagging. The study involved 780 respondents and used IBM SPSS 21.0, AMOS 21.0, Confirmatory Factor Analysis (CFA) and the Maximum Likelihood (ML) estimation method for data processing. The results suggested that

habitual and hedonic motivations have a substantial causal impact on the intentions of Facebook users to tag photos.

Allam et al. [28] focused on UTAUT2 which validated a THREE (3) dimensional hedonic motivation as major determinants of photo tagging. The online survey provided 187 responses, with 174 valid responses where PLS SEM approach was used for data analysis. The results indicated that users were more inspired by the sensations of explorability and enjoyment while using Facebook tagging tools.

## 3 Result and Discussion

As discussed in Sect. 2, the useful insights gained from the analysis can be used as techniques for attracting and retaining new users as well as generating and developing new visual knowledge that involves content and platforms. The strategies can enhance social media users' entertainment and enjoyment towards visual information. It is a significant influence to connect the content actions of users related to visual information with behaviors or technologies socially valued that could contribute to social recognition and validation as well as helping online social media users resolve psychological obstacles and promote self-reliance on their social media participation.

Moreover, the review may be able to improve managements, industries, business and marketers on visual information's uses to efficiently and effectively market products and communicate with its existing and potential clients. With the understanding of the visual information characteristics and benefits, companies which use visual advertisement on the social media can attract maximum number of customers which give the business a louder voice to promote their organization.

Finally, the review provides an understanding of why users utilize visual information in social media. It shows that visual information's uses via social media does transform the way people communicate and socialize. This review will be useful in designing a social media in a new way to ensure that users actively participate and engage in it on an ongoing basis. The media's own characteristics can therefore affect user effects, such as user involvement and participation in social media features; Comment, Like, Post, Reactions, Share, and Tags.

## 4 Conclusion

The studies show that Uses and Gratifications Theory (UGT) is the most accepted social media concept. In addition, UGT is widely used to explore and investigate the usage of social media motivations or gratifications. UGT appears to be the most significant literature motivations for social media and is known as a socio-psychological approach that is considered empirically positive. The theory presumes that individuals are engaged, reasonable and analytical in their decisions where, by

defining the rewards or benefits, they try to address the question as to why committed and rational users use the various aspects and features of social media.

The studies also show that audience reactions, enjoyment, entertainment and social interaction are the most significant factors contributing to user's intentions and behaviors towards contents, to engage and continue to use. Compared to other platform, more online social media users tend to create content and give feedback via Facebook. Images is the chosen visual information as it is the most popular content used. PLS SEM is the method used for data analysis, since the measuring properties of the constructs are less restrictive.

## 5 Future Works

Another line of work that may be pursued from this study would be on integrated technology adoption, demographics and other visual information to create a holistic understanding on how information in social media influences the users' intention and behavior. Future research can be performed on more important factors in improving the adoption model's predictability to better illustrate the user's visual information intentions. Additionally, the adoption model developed can be verified through data collection and analysis on empirical data.

## Appendix

Table 2.

**Table 2** Technology adoption model used in users' online social media behaviors

| Model | Factors | Media | Information | Intention | Analysis | References |
|---|---|---|---|---|---|---|
| Social Cognitive Theory (SCT) | Expected outcomes (social presence, social status), self-efficacy | Facebook | Photo | Tagging | SPSS 21.0. AMOS 21.0, CFA, ML | [22] |
| | Motivation (self-improvement, self enhancement, intention), model attractiveness, affective response (pleasant affection), self-efficacy | Instagram | Image | Post | ANOVA, Scheffee, CFA and EFA | [23] |

(continued)

**Table 2** (continued)

| Model | Factors | Media | Information | Intention | Analysis | References |
|---|---|---|---|---|---|---|
| Social Identity Theory (SIDT) | Cognitive, affective, evaluative | Instagram | Photo | Post | SPSS, CFA, SEM, AMOS, Interviews | [27] |
| Social Influence Theory (SIT) | Identification, internalization, compliance | Facebook | Not mention | Share | PLS SEM | [16] |
| Technology Acceptance Model (TAM) | Perceived usefulness, perceived ease of use | Instagram | Photo | Post | SPSS, CFA, SEM, AMOS, Interviews | [27] |
| | Perceived usefulness, perceived ease of use | Facebook | Photo | Tagging | PLS SEM | [28] |
| | Perceived usefulness, perceived ease of use | Facebook | Not mention | Reaction | PLS-SEM. PLS-MGA | [30] |
| Theory of Reasoned Action (TRA) | Social norms, social, presence, perceive ease, perceive usefulness | Facebook | Not mention | Like | EFA, CCA, HRA | [37] |
| Theory of Planned Behavior Model (TPB) | Attitude, subjective norm perceived behavioral control, narcissism | Facebook | Image | Post | HMRA | [14] |
| | Attitude toward behavior, subjective norm, perceived behavioral control, narcissism | Instagram | Photo | Post | AMOS 22 | [35] |
| Unified Theory of Acceptance and Use of Technology Version 2 (UTAUT2) | Habit, hedonic motivation, facilitating conditions, social influence, effort expectancy, performance expectancy | Facebook | Photo | Tagging | SPSS 21.0. AMOS 21.0, CFA, ML | [22] |
| | Hedonic motivation (curiosity, enjoyment, explorability) | Facebook | Photo | Tagging | PLS SEM | [28] |

(continued)

**Table 2** (continued)

| Model | Factors | Media | Information | Intention | Analysis | References |
|---|---|---|---|---|---|---|
| Uses and Gratifications (UGT) | Affection seeking, attention seeking, disclosure, entertainment, habitual pastime, information sharing, social influence, social interaction | Facebook | Photo | Share | EFA, MLE, SPSS 21.0 tool | [39] |
| | Audience, attractiveness, appropriateness, image quality/composition, subject, response, platform, privacy, online longevity, humor | Facebook | Image | Post | TA | [11] |
| | Content appreciation, friends, audience, reputation, appropriateness, support, do others 'like' it? | Facebook | Image | Like | TA | [11] |
| | Socialization, enjoyment, immersive experience, self-presentation, expression | Facebook | Not mention | Reaction | PLS-SEM. PLS-MGA | [30] |
| | Retention of moments, entertainment, expressive information sharing, social interaction, social use, habitual passing of time, relaxation, imaginary audience, social pressure and identity | Facebook | Photo | Post | MLRA | [6] |
| | Enjoyment, pleasing others, monetary incentive, passing time, interpersonal relationship | Facebook | Not mention | Like | EFA, HRA, CCA | [37] |

## References


1. Hlee S, Lee H, Koo C (2018) Hospitality and tourism online review research: a systematic analysis and heuristic-systematic model. Sustainability 10(4)
2. Taherdoost H (2018) A review of technology acceptance and adoption models and theories. Procedia Manuf 22:960–967
3. Edgley C (2018) Why your brain loves visual information—Ember Television. ember television. [Online]. Available: https://embertelevision.co.uk/blog/why-your-brain-loves-visual-information/. Accessed: 22-Apr-2020
4. Wang T, Lee FY (2020) Examining customer engagement and brand intimacy in social media context. J Retail Consum Serv 54(November 2019):102035
5. Hansen DL, Shneiderman B, Smith MA, Himelboim I (2020) Social media: new technologies of collaboration. Anal Soc Media Netw NodeXL, pp 11–29
6. de Vaate AJDNB, Veldhuis J, Alleva JM, Konijn EA, van Hugten CHM (2018) Show your best self(ie): an exploratory study on selfie-related motivations and behavior in emerging adulthood. Telemat Inf 35(5):1392–1407
7. Lee E, Lee JA, Moon JH, Sung Y (2015) Pictures speak louder than words: motivations for using Instagram. Cyberpsychol Behav Soc Netw 18(9):552–556
8. Simpson CC, Mazzeo SE (2017) Skinny is not enough: a content analysis of fitspiration on Pinterest. Health Commun 32(5):560–567
9. Grieve R (2017) Computers in human behavior unpacking the characteristics of Snapchat users: a preliminary investigation and an agenda for future research. Comput Hum Behav 74:130–138
10. Kwon SJ, Park E, Kim KJ (2014) What drives successful social networking services? A comparative analysis of user acceptance of Facebook and Twitter. Soc Sci J 51(4):534–544
11. Lowe-Calverley E, Grieve R (2018) Thumbs up: a thematic analysis of image-based posting and liking behaviour on social media. Telemat Inf 35(7):1900–1913
12. Shin J, Lee S (2020) Intimacy between actual users and virtual agents: interaction through 'likes' and 'comments,' In: Proc. 2020 14th International Conference on Ubiquitous Information Management and Communication IMCOM 2020, pp 1–4, 2020
13. Guy I, Ronen I, Zwerdling N, Zuyev-Grabovitch I, Jacovi M (2016) What is your organization 'like'? A study of liking activity in the enterprise. In: Conference on Human Factors in Computing Systems—Proceedings, pp 3025–3037
14. Lowe-calverley E, Grieve R (2018) Self-ie love: predictors of image editing intentions on Facebook. Telemat Inf 35(1):186–194
15. Elizabeth Stinson, Facebook Reactions, the Totally Redesigned Like Button, Is Here | WIRED. WIRED Mag 2016. [Online]. Available: https://www.wired.com/2016/02/facebook-reactions-totally-redesigned-like-button/. Accessed 05-Apr-2020
16. Oliveira T, Araujo B, Tam C (2020) Why do people share their travel experiences on social media? Tour Manag 78(December 2019):104041, 2020
17. Dhir A, Chen GM, Chen S (2017) Why do we tag photographs on Facebook? Proposing a new gratifications scale. New Media Soc 19(4):502–521
18. Hayes RA, Carr CT, Wohn DY (2016) One click, many meanings: interpreting paralinguistic digital affordances in social media. J Broadcast Electron Media 60(1):171–187
19. Bandura A (1999) A social cognitive theory of personality. Handb Personal Theory Res, pp 154–196
20. Muslim A, Harun A, Ismael D, Othman B (2020) Social media experience, attitude and behavioral intention towards umrah package among generation X and Y. Manag Sci Lett 10(1):1–12
21. Seear KH, Atkinson DN, Henderson-Yates LM, Lelievre MP, Marley JV (2020) Maboo wirriya, be healthy: community-directed development of an evidence-based diabetes prevention program for young Aboriginal people in a remote Australian town. Eval Program Plann 81(April):101818
22. Dhir A, Kaur P, Rajala R (2018) Why do young people tag photos on social networking sites? Explaining user intentions. Int J Inf Manage 38(1):117–127

23. Peng CT, Wu TY, Chen Y, Atkin DJ (2019) Comparing and modeling via social media: the social influences of fitspiration on male instagram users' work out intention. Comput Hum Behav 99(January):156–167
24. Tajfel H, Turner J (1979) An integrative theory of intergroup conflict. In: Austin WG, Worchel S (Eds) The social psychology of intergroup relations. Brooks/Cole, Monterey, CA, pp 71–112
25. Hong S, Jahng MR, Lee N, Wise KR Do you filter who you are? Excessive self-presentation, social cues, and user evaluations of Instagram selfies. Comput Hum Behav 104(October 2019):106159
26. Lee BK, Suh T, Sierra JJ (2020) Understanding the effects of physical images on viewers in social comparison contexts: a multi-study approach. J Promot Manag 26(1):1–18
27. Nedra BA, Hadhri W, Mezrani M (2019) Determinants of customers' intentions to use hedonic networks: the case of Instagram. J Retail Consum Serv 46(May 2018):21–32
28. Allam H, Bliemel M, Spiteri L, Blustein J, Ali-Hassan H (2019) Applying a multi-dimensional hedonic concept of intrinsic motivation on social tagging tools: a theoretical model and empirical validation. Int J Inf Manage 45(January):211–222
29. John SP, De'Villiers R (2020) Elaboration of marketing communication through visual media: an empirical analysis. J Retail Consum Serv 54(December 2019):102052
30. Shao C, Kwon, KH (2018) Clicks intended: an integrated model for nuanced social feedback system uses on Facebook. Telemat Inf (May):1–14, 2018
31. Bazi S, Filieri R, Gorton M (2020) Customers' motivation to engage with luxury brands on social media. J Bus Res 112(March):223–235
32. Pan Y, Torres IM, Zúñiga MÁ, Fazli-Salehi R (2020) Social network advertising: the moderating role of processing fluency, need for cognition, expertise, and gender. J Internet Commer 19(3):298–323
33. Ajzen I (1985) From intentions to actions: a theory of planned behavior. Action Control pp 11–39
34. Rubenking B (2019) Emotion, attitudes, norms and sources: exploring sharing intent of disgusting online videos. Comput Hum Behav 96(October 2018):63–71
35. Kim E, Lee JA, Sung Y, Choi SM (2016) Predicting selfie-posting behavior on social networking sites: an extension of theory of planned behavior. Comput Hum Behav 62:116–123
36. Abir T, Muhammad D, Yazdani NA, Bakar A, Hamid A, Survey H (2020) Electronic word of mouth (e-WOM) and consumers' purchase decisions: Evidences from Bangladesh. J Xi'an Univ Archit Technol XII(III):2004–2011
37. Lee SY, Hansen SS, Lee JK (2016) What makes us click like on Facebook? Examining psychological, technological, and motivational factors on virtual endorsement. Comput Commun 73:332–341
38. Schaffer DR, Debb SM (2020) Assessing Instagram use across cultures: a confirmatory factor analysis. Cyberpsychol Behav Soc Netw 23(2):100–106
39. Malik A, Dhir A, Nieminen M (2016) Uses and gratifications of digital photo sharing on Facebook. Telemat Inf 33(1):129–138

# A Pedagogical Framework with Integration of TPACK for Mobile Interactive System in Teaching Mathematics


**Daniel Lai, Lew Sook Ling, and Ooi Shih Yin**

**Abstract** Although there is a variety of technology available in 21st century, the way classes and lesson being conducted are still mostly remaining the same which is teaching via one-way communication. One-way communication teaching process surfastces issues like lack of interaction where minimal discussion is going on during teaching session, Since the interaction between teachers and students are fairly poor, limited classroom activity can be expected such as "Question and Answer" which lead to the increment of boredom in classes and lessons among students. As the teaching process is leaning towards instructional, students' feedback is usually being overlooked. Hence, since the availability of educational technology is getting more common nowadays, the integration of technology in classroom is encouraged changing the teaching and learning environment including knowledge delivery method from teacher perspective. Introducing mobile interactive system allows teachers deliver their knowledge differently, however, they are required to adapt and get familiar with the educational technology for improving teaching experience. Therefore, Technological Pedagogical Content Knowledge (TPACK) framework is applied in this paper to assess teachers using the seven elements of the framework. This paper is aimed to address the issues faced by conventional classroom and identify the effectiveness of teachers conducting classes using educational technology with the application of TPACK framework. Proposed TPACK framework is formed with the integration of three elements which are teachers' efficiency, students' performance and students' engagement representing the outcome of current TPACK framework.

D. Lai · L. S. Ling (✉) · O. S. Yin
Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia
e-mail: sllew@mmu.edu.my

D. Lai
e-mail: daniellai9317@gmail.com

O. S. Yin
e-mail: syooi@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_3

**Keywords** Teacher and student engagement · Mobile interactive system · Teaching and learning environment · Pedagogical framework · Technological pedagogical content knowledge (TPACK) framework

## 1 Introduction

The conventional classroom is undergoing major and rapid changes from past till now. Many technologies are invented throughout these years in various sectors including education sector. Due to the availability of technology nowadays, students are more preferred to stay connected with each other using technology [1]. Moreover, with the assistance of modern technology enabled work to be done more efficiently, effectively, convenient and most importantly it makes thing easier [2]. However, the way teachers and students used to communicate remained one-way communication where it leads to the lack of students' engagement in the classroom. The lack of student engagement leads to the drop of courses especially online courses that involved far less interaction between teachers and students [3]. Furthermore, the closeness of relationship between teachers and students plays an important role as the closer their relationship the better academic outcomes and self-esteem from students can be achieved [4]. Additionally, one-way communication serves two purposes which are sharing information and as a reminder of teachers' existence. Generally, subject like mathematics practices conventional way of teaching where teachers are dominant over students [5]. Besides, students are disallowed to express their thoughts and opinions because teachers wanted to gain full-control of the classroom [6]. Hence, gaining feedback from students' end is difficult. Without students' feedback, teachers might face difficulties in planning and preparing teaching materials that in-line with students' needs and capabilities.

Mobile application is introduced and accepted by majority of the society as it is a very commonly owned technological device. By integrating mobile technology into classroom, students can establish better communication with classmates and teachers [7]. Moreover, with the aids of technology in blended learning can boost up student engagement, enjoyment as well as academic achievement [8]. Nowadays, the interaction between human and computer had gone through tremendous changes due to the influence of computational devices that supporting touch [9]. Likewise, mobile devices are capable to be used as writing tools and incorporate into classroom [9]. Therefore, on-screen writing pad is introduced granting teachers with better mobility and flexibility in teaching with screen sharing.

### *1.1 Engagement Between Teachers and Students*

In this paper, the engagement between teachers and students is being studied and related to recent education changes. In England, during 19th century, the education

system neglected character-building quality and not creating intellectual curiosity [10]. In fact, having joy in the classroom was prohibited and integration of teacher's personalities was forbidden [11]. Hence, the enjoyment of learning cannot be achieved due to minimal interaction in the classroom. With the evolution of technology, the possibility of integrating technology into education sector is promising. According to Malik, the core changes brought by ICT in society has called for research on specific new forms of learning and epistemological issues regarding how learning occurs and how knowledge emerges beyond the borders of traditional systems of education [12]. The availability of these technologies allows teaching and learning becoming more interactive and interesting. According to Hussin, Generation-Z who are revolutionised by technology happened to act positively towards challenges, prefer group discussion and interactive learning environment [13]. Furthermore, the learning materials are required to be synchronised with the needs of learning process for better understanding regarding the concepts of subjects [14]. Moreover, teachers are playing a vital role as they will affect students' desire and developing their self-learning capability directly [15]. Additionally, student who spend more time in schooling tends to have lower dropout rate [16]. This is why John Dewey, an Educational Philosopher, mentioned more than a century ago, "If we teach today as we taught yesterday, we rob our children of tomorrow" [17]. Furthermore, technology is known to be the catalyst in communication, collaboration, innovation and problem solving in 21st century. Hence, the introduction of mobile interactive system as educational technology into schools and universities is imperative. It has the potential to be implemented in class activities increasing the interactivity between teachers and students [18]. Furthermore, mobile application also helps in refining and improving students' academic achievement as well as their learning capability [19]. It boosts up students' curiosity in terms of knowledge seeking makes them more attracted and interested in classroom activity. [20]. In fact, technology has the capability to increase student engagement as well as their understanding level towards learning content [21]. Through screen sharing, students can see the screen mirroring from teacher's computer screen where live-collaboration is happening at the same time [22]. The interactivity is improved because teachers are no longer required to reach out whiteboard or blackboard back and forth illustrating their ideas to students. Screen sharing is able to helps teachers showing their writing from on-screen writing pad to students in synchronised manner with the writing content in the mobile interactive system without sacrificing teachers' mobility.

Teacher is known to be a professional who responsible for delivering knowledge to students, however, the knowledge sometimes can be forgotten over time. Students tend to be more interested in lesson taught by their favourite teacher than the one that they dislike. Along with teaching methods, the image of the teacher attaches great importance in the process of forming knowledge [23]. From the perspective of students, it is very clear that teacher plays an important role in their learning process. Besides, teacher is also known as a leader in guiding students throughout lessons [24]. As a teacher, he or she does not only require to teach students but also understand their behaviour and desire as well as improving their learning ability and capability [25]. Therefore, teacher's qualities are vital when attempting to bond with

students as teacher's attributes shaped his or her image reflecting how students see teacher from their perspective.

According to Dakhane, the presence and popularity of mobile devices and mobile applications are common amongst students [26]. Education sector is assisted by various educational technologies to improve student engagement in classes especially in higher education. It brings positive impact in affecting student's experience in learning. Moreover, conventional teaching method is ineffective when it comes to student engagement as teacher can hardly focusing on students where teachers' attention is constantly diverted. In conventional classroom, teacher is the only active person who control the entire classroom activities and expecting students concentrate in classroom but it proven otherwise [27]. With the help of technological devices, teachers are given better flexibility and mobility. High flexibility and mobility allow teachers to pay more attention and interact with students effectively. Putra mentioned that teaching and learning process is an interaction between teachers and students which are two different things but form a unity [28]. Compared to conventional method, teaching and learning are conducted differently these days. Teaching and learning can be conducted through mobile devices, online, augmented reality, virtual reality and other state-of-art technology [29]. These technologies create new possibilities for teacher implementing a variety of methods turning teaching and learning environment become more interaction-oriented. It allows teachers have better interaction with students by using educational technology. Educational technology boosts students' curiosity driving students to further explore and gaining interest in learning.

## *1.2 Technological Pedagogical Content Knowledge Theoretical Framework*

Pedagogical framework is a structure created based on the philosophy of teaching and learning. It serves as guidelines for teacher to do evaluation and refinement when they facilitate classes. This set of guidelines is rather well-known among colleges and universities for its consistency. The pedagogical framework is meant to be a supportive medium to assist teacher delivering and transforming knowledge to students maintaining a high-quality content for both teaching and learning based on best practices [30]. With the help of pedagogical framework, teachers are able to address students with the most effective way of absorbing the knowledge being taught. According to Fairholme College, the pedagogical framework has to be coordinated with a few principles which are safe, supportive, connected and inclusive learning environment, student-centered planning, evidence-based decision making, targeted and scaffolded instruction, alignment of curriculum pedagogy and assessment as well as high expectation [31]. According to Fazidah Naziri, teachers need to equip themselves with a transformation in knowledge and technology to educate Z generation today [32]. Besides, Utami et al. also mentioned that technological involvement in learning activities cannot be avoided as technology integration in

teaching is considered as a more effective approach [33]. The concept of pedagogical content knowledge theoretical framework, TPACK reflects the status of technological, pedagogical, and content knowledge of educators [34]. Based on TPACK model, lecturers' knowledge about technology and experience the successful integration of collaborative tools in teaching and learning environment are the keys for effective adoption of new technologies [35]. The combination of knowledge and pedagogy with the involvement of information technology enabled teachers in developing pedagogical knowledge, skills and enhancing students' learning [36]. Student learning capability will be improved presuming teacher himself or herself is willing to refine their technological knowledge and integrate it into their content knowledge and pedagogical knowledge [37]. Teachers are expected to demonstrate how technology can be implemented in supporting the learning content, how specific pedagogies best support with the use of and facilitate learning. Additionally, TPACK is also reflecting the interdependence of three contributing knowledge domains which are Content Knowledge (CK), Pedagogical Knowledge (PK), and Technology Knowledge (TK) in a better picture [38]. From teacher perspective, mastering TPACK from the aspects of knowledge, skills and technology is necessary. Being innovative in teaching allows teachers to be in line with up-to-date technological world in educating the next generation.

### *1.3 Mobile Interactive System*

TPACK framework is applied in this study due to the substantial involvement of technology in teaching and learning. As advanced technologies also offer interactivity, these known technologies were focused by other studies [32–34, 37, 38] while this study is focusing on delivery method which is also referring as mobile interactive system. Solomon et al. mentioned that the specifications of current smartphones are well-equipped with high-end memory, processor, display and battery which are capable to serve learners, teachers and researchers with more possibilities [39]. Stathopoulou et al. also mentioned that mobile devices are having the potential to alter the classroom from conventional to a more interactive and engaging [40]. Due to the massive and hasty growth of the usage of mobile technology in campuses, learning activities and research will be depending on mobile technology in future classroom [41]. In this study, mobile interactive system is a mobile application that emphasises on high mobility and low hardware requirement where expensive or external equipment are not required. It can mitigate one-way communication between teachers and students. It also can change teachers' teaching experiences from various aspects. The most spectacular change is the teaching environment where two-way communication is established instead of one-way communication. Teachers can maintain interaction with students by making use of other existing technologies with features provided by the mobile interactive system simultaneously. Moreover, teachers and students can bring it to any classes that require the use of it without any hassle. Therefore, the objectives of this study are as follows:

1. Identify current trend in education sector which addresses the conventional classroom issues.
2. Identify the effectiveness of teachers conducting classes using educational technology with the application of TPACK framework.

## 2 Literature Review

In this paper, the literature review is being done based on two major aspects which are interactive learning and on-screen writing pad with screen sharing. These two aspects are able to illustrate interactivity better as interactive whiteboard is already being introduced as educational technology and implemented in classroom [42]. Hence, in-depth understanding about these two aspects is essential to cater with current trend of educational technology. As mentioned earlier, having more interactivity in the classroom tend to result in better engagement between teachers and students [43–45]. Thus, interactive learning is one of the key aspects involving how teachers apply technology changing the teaching and learning environment. Similar to interactive whiteboard, on-screen writing pad with screen sharing serves teachers with lower cost and higher portability [46]. This study aimed to mitigate one-way communication teaching in classroom by introducing more instructiveness into classroom activities.

### *2.1 Mobile Interactive System*

Sahronih et al. mentioned that interactive learning media is having the capability of describing teachers' message to student forming a two-way communication [47]. Hence, interactive learning has a strong bond with educational technology. The use of the interactive learning method alongside with educational technology creates more possibilities in learning process which motivates students and creates less confusion and difficulty in both teaching and learning process [48]. According to Wang et al., traditional personal computer will be replaced by mobile devices slowly within the next few years and becoming the main learning equipment especially in tertiary education [49]. Educational technology is a toolkit that creates educational contents with appropriate technology making an interactive learning environment. Therefore, students should not be passive in exploring and accepting new information. It is vital for classroom studies to be carried out in an interactive environment as it helps learner developing their independence.

Oluwajana et al. stated that students' active participation in learning process is achievable via interactive learning [50]. Interactive learning method consists of two types of interaction which are cooperation and competition, rivalry [51]. Alexandrovna mentioned that cooperation is meant for promoting teamwork and aiming for achieving one goal whereas competition, rivalry is meant for opposition of goal and opinions [51]. The implementation of interactive learning helps students to be more

engaged in classroom, improves their problem-solving skills and critical thinking skills. Furthermore, the dependence of oneself towards smartphone has becoming a norm where online learning is doable other than serving solely as a communication platform [52]. In contrast with Multimedia Interactive Learning Online (MILO) suggested in the research of Pakyuan et al., proposed mobile interactive for this study is focusing on delivery method rather than content itself. However, both of the studies share similar aim which is maintaining the function of teacher in learning process [52]. Nina and Heru also mentioned that the innovation in mobile devices making smartphone ideal for both learning and education [53]. They also mentioned that the difficulties like formulas and calculations in learning physics can be resolved with the use of technology because it helps to improve learners in understanding diagrammatic and argumentative representation with better efficiency [53].

## *2.2 On-Screen Writing Pad with Screen Sharing*

Recently, interactive whiteboard (IWB) has been a trend in education sector where IWB serves as an educational tool [54, 55]. It is commonly found in Western schools. The IWB is a large touch screen with the capability of replacing the conventional setup in classroom. It enhances teaching environment by introducing interactivity and creating visual impact in the classroom. Tsai (2019) mentioned that the difficulties faced by teachers in using interactive whiteboard are the lack of computer competency and insufficient technical support [56]. Moreover, further training for teachers is needed as the technical support is insufficient and ineffective [57]. Moreover, the implementation of IWB in conventional classroom does not lead to negative effect to the teaching method which favored by most teachers.

Comparing IWB to the on-screen writing pad introduced in this study, from the perspective of hardware, IWB is more hardware dependent as it requires the use of special made whiteboard to be installed and configured whereas on-screen writing pad is less dependent on hardware as teachers only require to operate the system using Android smartphone and a computer for server hosting. According to Andy (2018), the utilisation of IWB's main purpose, interactivity for teaching and learning is ambiguous. In the meantime, since IWB has smaller size compared to traditional projection screen, the proposed mobile interactive system has an edge because its screen sharing feature can be projected through the ordinary projection screen without sacrificing the mobility of teachers as the proposed mobile interactive system can control the connected computer from distance [58]. Additionally, IWB requires teachers' prior knowledge in designing teaching plans and materials based on IWB capabilities to allow students interact with IWB and learning materials [59]. The difference between IWB and proposed mobile interactive system from previous context is the ease of use of the technology. As proposed mobile interactive system is an Android application installed in the smartphone, the learning curve for teacher to use the system is relatively low compared to IWB.

# 3 Methodology

In this paper, a systematic literature review (SLR) is being carried out with the use of Multimedia University Library Integrated Access (MULIA) 3.0 and Google Scholar in getting relevant sources for the study. The relevant sources are being searched based on keywords. This paper adopted the search procedure mentioned in [60]. The research questions of this paper are RQ1_How mobile interactive system enhances teachers and students' engagement in classroom? and RQ2_How TPACK framework affects teaching and learning environment? Firstly, keywords are being formulated based on research questions including teacher and student engagement, mobile interactive system, teaching and learning environment and TPACK theoretical framework. These words are then undergoing a process called Systematic Search Procedure which is stated in [60]. In Systematic Search Procedure, there are several substages where it begins with a set of search words which are obtained from thesaurus based on formulated keywords for semantic criteria search. The semantic structure of the paper title is taken into consideration when searching for specific papers. Next, making use of appropriate search script by referring to syntax of various databases such as Scopus, Google Scholar and other similar databases. With proper search script applied, relevant papers are then selected from databases relatively. At last, a list of papers as search results are discovered to be further reviewed. The process of systematic search procedure will be continued until research questions are being answered.

## *3.1 Systematic Literature Review*

Systematic literature review is also known as SLR in which research questions can be identified. Additionally, SLR is also crucial for the justification of future research for that sector. SLR is important to those who are newly started in their specified fields to learn and explore for up-to-date information and work which is related to the field of study [60]. The information gathered includes the method that has been implemented and results which are beneficial for students having more insights about the area of interest in terms of study. Information that is commonly acquired via SLR is the databases of the work, publications of work presented and research centres of certain work. The gathered information is presented statistically. Since researchers are not always familiar with the fields of study, they usually face circumstances such as the lack of related knowledge and identification of journals. From the perspective of researchers from similar research fields, they might share their findings in addressing or resolving similar problems as well as objectives. Their findings include their progress, result, approaches, networks and so forth. As a researcher, interaction and contribution are involved when conducting research study. Difficulties and challenges such as key words of search, inclusion and exclusion criteria for filtering search results will be faced due to the unfamiliarity of certain research fields. Therefore, mentefacto conceptual is one of the mandatory tools that leads to good reading and learning [60].

In this study, four questions are required to be answered which are what characterises it, in essence? In what group of things include it? What are your differences with similar objects? and, are there subtypes of yours? By answering these questions, four groups of thoughts are formed resulting in isoordinated, superordinated, excluded and infraordinated.

## 4 Expected Outcome

### *4.1 Proposed TPACK Framework*

TPACK is also known as technological pedagogical and content knowledge which require teachers to have a thorough understanding about the knowledge, skills and technology so as to improve teachers' efficiency in teaching. In 21st century, teachers are required to adapt to the latest technological trends in order to be accord with 4th Industrial Revolution. In such revolution, Information and Communication Technology is becoming a very important medium in between teachers and students. The TPACK framework was being introduced and enhanced by Mishra and Koehler in year 2006. It is provided for educators a more rounded experience of teaching [61]. In TPACK, it consists three major elements, technological knowledge, content knowledge and pedagogical knowledge which symbolise the combination of knowledge and pedagogy with the implementation of information technology. A total of seven elements can be found in TPACK framework. The major elements of TPACK consist of three major area of knowledge which are technological knowledge (TK), pedagogical knowledge (PK) and content knowledge (CK). TK is the knowledge or the cognition of various technologies regardless the tier of those technologies to improve teaching efficiency. PK is thorough knowledge of process or method of teaching and learning which became the value of education to improve students' engagement in classroom, CK is the subject that needed to be taught and learned to enhanced students' performance academically, Hence, the interrelation element of TK and PK is TPK which is also known as technological pedagogical knowledge. TPK is the knowledge of pedagogical activities of teachers with proper understanding of technology implementation that is best suited for desired teaching and learning experiences. TPK is referring to teachers' understanding in making use of mobile interactive system and its features to assist them in changing teaching and learning experiences according to their preferences. Also, the interrelation element of TK and CK is TCK which is also stands for technological content knowledge. TCK is the knowledge of representing the concept of technology into a subject. It is referring to teachers' understanding on how the subjects can be best represented and influenced by technology which is referring to mobile interactive system with on-screen writing in this study. Pedagogical content knowledge (PCK) is the interrelation element of PK and CK. It is teachers' ability in modifying their content knowledge to fit in teaching approach appropriately. It is also referring to teachers' understanding in choosing the

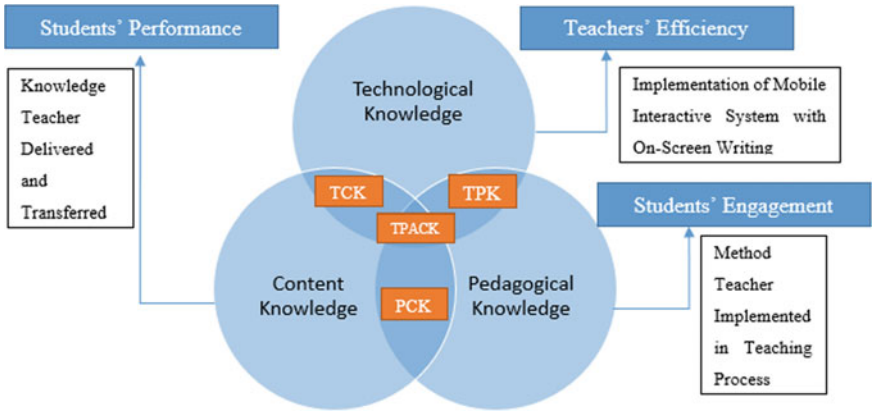



**Fig. 1** Proposed TPACK framework

appropriate method in delivering CK in different level in terms of grade as teachers; teaching practices can be improved with the assistance of thorough understanding in PCK. As for TPACK, it is the combination of all three knowledge which utilises the state-of-art technology that can cope with the needs and wants of students.

In this study, the proposed TPACK framework in Fig. 1 consists of additional elements which are teacher's efficiency, students' performance and students' engagement. Teacher's efficiency in this study is referring to the efficiency of teacher using technology in the teaching process. Therefore, it involves elements that related to TK which are TK, TCK, TPK and TPACK. The level of teachers' understanding about technology will be assess and reflects how efficient are they when lessons are taught with the integration of technology in classroom activities. As for students' performance, it is assessed with CK related elements. CK related elements include CK, TCK, PCK and TPACK. The relevancy of teaching materials prepared by teachers will affect students' performance either positively or negatively. As for students' engagement, PK related elements such as PK, PCK, TPK and TPACK will be taken into consideration for assessment. The interactivity and effectiveness of teachers' delivery method for the subjects he or she taught will be recorded and serves as input for determining students' engagement in the classroom.

## 5 Conclusion

The current paper proposes a mobile interactive system for learning by integrating pedagogical framework with technology. The research questions of this study are being identified through SLR whereby justification for the future research of current study can be made. The system is intended to eliminate conventional teaching method which is conducting through one-way communication which resulting in less student

engagement and interaction.in classes. Two-way communication is proposed by mobile interactive system to address classroom communication issue whereby two-way communication is more effective in teaching as it involves both parties which are referring to teachers and students in this context. Having a more effective communication in classroom leads to higher student engagement. Besides educational technology, which is referring to proposed mobile interactive system, teachers are also required to come out with a well-prepared teaching plan and teaching materials with the help of TPACK. The proposed TPACK framework is having the intention allowing teachers to tailor their teaching content according to students' performance and capability, one should also have in-depth understanding about the content, knowledge and specific pedagogies as well as knowing how to use the latest technology and integrate the technology into teaching. Without prior knowledge in any one of the stated elements which are referring to pedagogical knowledge, technological knowledge and content knowledge, teachers' teaching experience and performance will be affected.

**Acknowledgements** This research was funded and supported by Multimedia University and Fundamental Research Grant Scheme (FRGS), Malaysia.

## References

1. Hoffmann MM, Ramirez AYF (2018) Students' attitudes toward teacher use of technology in classrooms. In: Multicultural education. Caddo Gap Press. LNCS, vol 9999, pp 1–13. Springer, Heidelberg
2. Wang LYK, Lew SL, Lau SH, Leow MC (2019) Usability factors predicting continuance of intention to use cloud e-learning application. Heliyon 5(6). https://doi.org/10.1016/j.heliyon.2019.e01788
3. Hussain M, Zhu W, Zhang W, Abidi SMR (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput Intell Neurosci https://doi.org/10.1155/2018/6347186
4. Xerri MJ, Radford K, Shacklock K (2018) Student engagement in academic activities: a social support perspective. High Educ 75(4):589–605. https://doi.org/10.1007/s10734-017-0162-9
5. Milenković A, Dimitrijević S (2019) Advantages and disadvantages of heuristic teaching in relation to traditional teaching-the case of the parallelogram area. European Mathematics Society. https://dms.rs/wp-content/uploads/2019/12/Zbornik-ERME.pdf#page=75
6. Otukile-Mongwaketse M (2018) Teacher centered dominated approaches: their implications for todays inclusive classrooms. Int J Psychol Counselling 10(2):11–21. https://doi.org/10.5897/ijpc2016.0393
7. Dobbins C, Denton P (2017) MyWallMate: an investigation into the use of mobile technology in enhancing student engagement. TechTrends 61(6):541–549. https://doi.org/10.1007/s11528-017-0188-y
8. Morris NP, Lambe J, Ciccone J, Swinnerton B (2016) Mobile technology: students perceived benefits of apps for learning neuroanatomy. J Comput Assist Learn 32(5):430–442. https://doi.org/10.1111/jcal.12144
9. Ebrahim AA, Tawfik AE (2018) Mobile applications for education. Int J Learn Teach 251–254. https://doi.org/10.18178/ijlt.4.3.251-254

10. Dey S (2019). Elementary education in 19th-century Bengal education policy of the British. Econ Polit Weekly 54(30):37–44.2
11. Smith V, Wortley A (2017) "Everyone's a comedian". No really, they are: using humor in the online and traditional classroom. J Instruct Res 6(1). https://doi.org/10.9743/jir.2017.3
12. Malik RS (2018) Educational challenges in 21st century and sustainable development. J Sustain Dev Educ Res 2(1):9. https://doi.org/10.17509/jsder.v2i1.12266
13. Hussin A (2018) Education 4.0 made simple: ideas for teaching. Int J Educ Literacy Stud 6(92). https://doi.org/10.7575/aiac.ijels.v.6n.3p.92
14. Krishnasamy S, Ling LS, Kim TC (2020) Improving learning experience of probability and statistics using multimedia system. Int J Emerg Technol Learn 15(1):77–87. https://doi.org/10.3991/ijet.v15i01.11349
15. Guerra A, Spliid CM (2018) Pedagogical development for course change and academic staff expectations. In: Engineering 2030: conceptualization of Industry 4.0 and its implications for engineering education, 7th edn. pp 572–582. Aalborg Universitetsforlag. https://vbn.aau.dk/ws/portalfiles/portal/302786154/2018_IRSPBL_Proceedings_Innovation_PBL_and_Competence.pdf#page=562
16. Zerihun Z, Kassahun A, Wassie C, Ebrie S, Rebso M (2019) Students' academic performance in conventional and alternative schooling: field based evidence. Eur J Altern Educ Stud 4(2). http://doi.org/10.5281/zenodo.3572787
17. Dewey J (1916) Democracy and Education. Free Press, New York http://www.ilt.columbia.edu/publications/dewey.html. Accessed 21 Nov 2005
18. Çimşir BT, Uzunboylu H (2019) Awareness training for sustainable development: development, implementation and evaluation of a mobile application. Sustainability (Switzerland) 11(3). https://doi.org/10.3390/su11030611
19. Etcuban JO, Pantinople LD (2018) The effects of mobile application in teaching high school mathematics. Int Electron J Math Educ 13(3). https://doi.org/10.12973/iejme/3906
20. Fabian K, Topping KJ, Barron IG (2018) Using mobile technologies for mathematics: effects on student attitudes and achievement. Educ Technol Res Dev 66(5):1119–1139. https://doi.org/10.1007/s11423-018-9580-3
21. Khan M, JeyaGopi R (2019) Evaluating readiness in using mobile devices for educational purposes among educators and students. In: International conference on innovation and technopreneurship 2019, vol 2019, no 007. http://eprints.intimal.edu.my/1280/1/vol.2019_007.pdf
22. Sarwate A, Tsuchiya T, Freeman J (2018) Collaborative coding with music: two case studies with EarSketch. In: Monschke J, Guttandin C, Schnell N, Jenkinson T, Schaedler J (eds.) Proceedings of the international web audio conference. TU Berlin. https://www.youtube.com/watch?v=0qBVSCRpggg
23. Gulnar A (2018) Modern teacher role for increasing the students' competence in pedagogical specialty. Opcion
24. Warren LL (2018) Behaviors of teacher leaders in the classroom. Psychol Behav Sci 7(6):104–108. https://doi.org/10.11648/j.pbs.20180706.12
25. Bih Ni L, Rabe Z, Asyikin Hassan N (2018) World academy of science, engineering and technology. Int J Educ Pedagogical Sci Teach Leadersh Dimension Hist Learn 12. https://www.researchgate.net/publication/329771175_Teachers_Leadership_Dimension_in_History_Learning
26. Dekhane S, Xu X, Tsoi MY (2013) Mobile app development to increase student engagement and problem solving skills. J Inf Syst Educ 24(4):299–308
27. Zaheer Abdul Ghafoor YA (2018) Integrating MALL into english flipped classroom at the tertiary level. India's High Educ Authority UGC Approv List J Ser Number 18(2):98–108. www.languageinindia.com
28. Putra CA (2018) Enhanced learning outcomes using interactive edutainment learning method. In: Proceedings international conference BKSPTIS 2018
29. Rahman MA, Ling LS, Yin OS (2020). Augmented reality for learning calculus: a research framework of interactive learning system. In: Lecture notes in electrical engineering. Springer, vol 603, pp 491–499. https://doi.org/10.1007/978-981-15-0058-9_47

30. Raju B (2020) Innovations and best practices towards quality education. Aayushi Int Interdiscip Res J (AIIRJ)., 40(50):2258–2262. https://archives.tpnsindia.org/index.php/sipn/article/view/3710/3593
31. Fairholme College Fairholme College Pedagogical Framework of Learning and Teaching (2018). https://www.fairholme.qld.edu.au/module/documents/download/1537
32. Naziri F, Rasul MS, Affandi HM (2019) Importance of technological pedagogical and content knowledge (TPACK) in design and technology subject. Int J Acad Res Bus Soc Sci 9(1). https://doi.org/10.6007/ijarbss/v9-i1/5366
33. Utami P, Pahlevi FR, Santoso D, Fajaryati N, Destiana B, Ismail ME (2019) Android-based applications on teaching skills based on TPACK analysis. In: IOP conference series: materials science and engineering. https://doi.org/10.1088/1757-899X/535/1/012009
34. Rahman SMM, Krishnan VJ, Kapila V (2017) Exploring the dynamic nature of TPACK framework in teaching STEM using robotics in middle school classrooms. In: ASEE annual conference and exposition, conference proceedings. https://doi.org/10.18260/1-2-28336
35. Vasodavan V, Dewitt D, Alias N (2019) TPACK in higher education: analysis of the collaborative tools used by lecturers. Asia Pac J Curriculum Teach 7:9–17
36. Malik S, Rohendi D, Widiaty I (2019) Technological Pedagogical Content Knowledge (TPACK) with information and communication technology (ICT) integration: a literature review. Atlantis Press. https://doi.org/10.2991/ictvet-18.2019.114
37. Karns SJ (2019) Pairing a learning activity types short course with collaborative curriculum design: an approach to impact teacher's technological pedagogical content knowledge. Cyberlaw Global E-business. https://doi.org/10.4018/978-1-59904-828-4.chtpg
38. Nurhadi D, Purwaningsih E, Masjkur K, Nyan-Myau L (2019) Using TPACK to map teaching and learning skills for vocational high school teacher candidates in Indonesia. Atlantis Press. https://doi.org/10.2991/ictvet-18.2019.9
39. Oyelere SS, Suhonen J, Wajiga GM, Sutinen E (2018) Design, development, and evaluation of a mobile learning application for computing education. Educ Inf Technol 23(1):467–495. https://doi.org/10.1007/s10639-017-9613-2
40. Stathopoulou A, Karabatzaki Z, Tsiros D, Katsantoni S, Drigas A (2019) Mobile apps the educational solution for autistic students in secondary education. Int J Interact Mob Technol 13(2):89–101. https://doi.org/10.3991/ijim.v13i02.9896
41. Karabatzaki Z, Stathopoulou A, Kokkalia G, Dimitriou E, Loukeri PI, Economou A, Drigas A (2018) Mobile application tools for students in secondary education. An evaluation study. Int J Interact Mob Technol 12(2):142–161. https://doi.org/10.3991/ijim.v12i2.8158
42. Şengül M, Türel YK (2019) Teaching Turkish as a Foreign language with interactive whiteboards: a case study of multilingual learners. Technol Knowl Learn 24(1):101–115. https://doi.org/10.1007/s10758-017-9350-z
43. Zimmermann P, Stallings L, Pierce R, Largent D (2018) Classroom interaction redefined: multidisciplinary perspectives on moving beyond traditional classroom spaces to promote student engagement. J Learn Spaces 7(1):45–61
44. Kaewunruen S (2019) Enhancing railway engineering student engagement using interactive technology embedded with infotainment. Educ Sci 9(2). https://doi.org/10.3390/educsci9020136
45. Cronhjort M, Filipsson L, Weurlander M (2018) Improved engagement and learning in flipped-classroom calculus. Teach Math Appl 37(3):113–121. https://doi.org/10.1093/TEAMAT/HRX007
46. Kukulska-Hulme A, Viberg O (2018) Mobile collaborative language learning: state of the art. British J Educ Technol Blackwell Publishing Ltd. https://doi.org/10.1111/bjet.12580
47. Lew SL, Ooi SY, Muthukumar Y, Rahman A (2018) Improving interactivity via icontrol: a presentation mobile app. In: Proceedings of the European conference on e-learning, ECEL, vol 2018, pp 308–315. Academic Conferences Limited
48. Putra CA (2018) Enhanced learning outcomes using interactive edutainment learning method. In: Proceedings international conference Bksptis 2018, pp 219–224. http://jurnal.unissula.ac.id/index.php/bksptis/article/view/3576/2614

49. Ruihu W, Guoli Y (2019) Exploration and Practice of Mobile InteractiveInternent Classroom Teaching Based on UMU. EasyChair. https://wvvw.easychair.org/publications/preprint_open/qJh5
50. Dokun O, Nat C, Muesser, Samson F (2019) An investigation of student's interactivity in the classroom and interactivity within learning management system to improve learning outcomes. Croatian J Educ—Hrvatski časopis za odgoj i obrazovanje 21. https://doi.org/10.15516/cje.v21i1.3085
51. Alexandrovna RN (2018) Development of Independence among future primary school teachers by applying interactive learning methods. J Educ E-Learn Res 5(2):118–121 https://doi.org/10.20448/journal.509.2018.52.118.121
52. Woo P, Shahril AM, Azmi E, Rosli H (2019) Interactive learning online: a case study of front office teaching and learning in higher learning institution in Malaysia. https://doi.org/10.6007/ijarbss/v8-i15/5101
53. Liliarti N, Kuswanto H (2018) Improving the competence of diagrammatic and argumentative representation in physics through android-based mobile learning application. Int J Instruct 11(3):106–122. https://doi.org/10.12973/iji.2018.1138a
54. Kutluca T, Yalman M, Tum A (2019) Use of interactive whiteboard in teaching mathematics for sustainability and its effect on the role of teacher. Discourse Commun Sustain Educ 10(1):113–132. https://doi.org/10.2478/dcse-2019-0009
55. Olivares DD, Castillo RR (2018) ICT in the classroom: primary education student teachers' perceptions of the interactive whiteboard during the teaching practicum. Educ Inf Technol 23 https://doi.org/10.1007/s10639-018-9716-4
56. Tsai C-C (2019) A study of taiwanese elementary school english as a foreign language: teachers' beliefs, advantages, and difficulties of using interactive whiteboards. Asia-Pac Soc Sci Rev 19:87–99
57. Benoit A (2018) Investigating the impact of interactive whiteboards in higher education: a case study. J Learn Spaces 7(1):76–90
58. Ahmad W, Ali Z (2019) Interactive Whiteboard (IWB) effectiveness in vocabulary achievement and motivation: Saudi EFL learners' perceptions and insights. Bull Adv Eng Stud 2(2):90–103. https://doi.org/10.31559/baes2019.2.2.5
59. Olivares DD, Castillo RR (2018) ICT in the classroom: primary education student teachers' perceptions of the interactive whiteboard during the teaching practicum. Educ Inf Technol 23(6):2309–2321. https://doi.org/10.1007/s10639-018-9716-4
60. Torres-Carrion PV, Gonzalez-Gonzalez CS, Aciar S, Rodriguez-Morales G (2018) Methodology for systematic literature review applied to engineering and education. In: IEEE global engineering education conference, EDUCON, vol 2018, pp 1364–1373. IEEE Computer Society. https://doi.org/10.1109/EDUCON.2018.8363388
61. Adams C (2019) TPACK model: the ideal modern classroom. In: Technology and the curriculum: summer (2019). https://techandcurr2019.pressbooks.com/chapter/tpack-modern-classroom/

# Towards Palm Bunch Ripeness Classification Using Colour and Canny Edge Detection


**Ian K. T. Tan, Yue-Hng Lim, and Nyen-Ho Hon**

**Abstract** The ripeness of the farm-able palm fruits is an important factor in the production of quality palm oil. The work presented is an image processing implementation in the palm oil industry to eliminate human errors in the judgment of the ripeness of palm fruit bunches as well as to introduce automation. Various techniques were employed to obtain data from the images provided for the data mining process. The features used are the colour of the palm fruit bunches and the amount of edges representing visible leaves in the palm fruit bunches, indicating empty sockets. The project is able to achieve an accuracy of up to 79.11%.

**Keywords** Ripeness · Palm kernel · Colour detection · Canny edge · Empty sockets

## 1 Introduction

The determining factor of palm oil production starts with the classification of the palm fruit bunch ripeness. This classification process is typically done manually and is prone to human errors, availability of human experts, and inconsistencies in the classification process due to various environmental aspects such as lighting. This manual process depends on visual cues; such as colour, texture, and the shape of the oil palm fruit bunch. Furthermore, the speed of this classification is an important factor for consideration. The palm fruit bunches are delivered in batches (truckloads) which arrives at irregular intervals. The pressure to classify them quickly increases

I. K. T. Tan (✉)
Monash Industry Palm Oil Research Platform, Monash University Malaysia, School of IT, 47500 Subang Jaya, Selangor, Malaysia
e-mail: ian.tan1@monash.edu

Y.-H. Lim
Innov8tif Solutions Sdn Bhd, Subang Jaya, 47650 Selangor, Malaysia
e-mail: neil@innov8tif.com

N.-H. Hon
Melangking Oil Palm Plantation Sdn Bhd, Sandakan 90000, Sabah, Malaysia
e-mail: nyenho.hon@mopp.com.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_4

the amount of human errors. Hence the use of technology is needed to address these drawbacks.

Technology that assist in moving towards automation of this classification process is an important area as the yield and quality of the palm oil production is highly dependent on the ripeness of the palm fruit bunches.

## 2 Literature Review

Image processing for farming had been researched and implemented for actual use to increase yield as well as to introduce automation for error reduction. One of the earlier published work in this area was by Meyer et al. [1] in 1998. Their work was to automatically differentiate weeds from corn crops in order to implement an anti-weed strategy using computer vision and statistical analysis. Features such as texture and excess of the plant's green colour were used in their work.

In recent years, image processing to determine ripeness has also been attempted. Abbaszadeh et al. [2] used image processing on the rind texture of watermelons to determine their ripeness. The pattern on the rind is processed to determine the stretch of the wavy patterns which in turn provides the cue on the ripeness of the watermelon. In their work, they also used colour analysis to grade the ripeness appropriately.

In the specific area of palm fruit ripeness, Choong et al. [3] published that the oil content of palm fruits is highly correlated to the redness of the palm oil fruits. They used a controlled environment where the height of the camera to the fruit is always at the same exact height (which means that the camera has to move to compensate for the fruit size) and lighting is illuminated from all angles in order to capture the images. The images were then individually manually edited for consistency using image processing application.

Ghazali et al. [4] used the RGB (Red, Green and Blue) colour components as the features for their work. In their work, they processed the images captured by eliminating the non-red colours and used the resulting images to classify them into three categories; ripe, under-ripe and unripe. Their work was limited by the carefully curated 30 sample images for each of the classification categories. Their work claimed to have an accuracy of 100% of the ripe category, and between 80-85% for the under-ripe and unripe categories.

A more recent work by Shabdin et al. [5], conducted a similar research to use the colour components to determine the ripeness of the palm fruit bunches but they included the use of the hue saturation and the colour intensity as the main feature. For their analysis, they used Artificial Neural Networks (ANN) and they reported an overall accuracy of 70%.

Using similar techniques as Shabdin et al. [5], Saaed et al. [6] also conducted their classification using ANN but with specialized hyperspectral active sensor system. The equipment used has 824 spectral bands which covers the colour frequency range of 400 to 1000 nm. This range of orange and red are in the region of 590–625 and

625–740 nm respectively and hence the sensor they used is of a very high accuracy and can be configured to capture the colour ranges accurately.

Although Ghazali et al. [4] claimed an accuracy of 100% for ripe palm fruit bunches and a very good 80-85% accuracy for the other two categories, the images were carefully curated and does not represent the actual plantation environment. Even with additional features included by Shabdin et al. [5], the overall accuracy is about 70%.

## 3 Data

The image collection by past researchers tend to veer towards the highest quality images possible with minimal noise. Sophisticated capturing devices, such as Hyperspectral Sensor [6], Microsoft Lifecam NX-600 [7] and Vivotek IP8332 Network Bullet Cameras [8] were used in a highly controlled lighting environment in order to minimize noise, lighting differences and varying backgrounds. Past work with high accuracy rates were likely to have used images captured in a highly controlled environment to ensure high accuracy of classification.

In the work presented in this manuscript, the images used were captured by a camera phone at an actual palm fruit bunch sorting area and not in a laboratory. The use of a camera phone is to simulate the use of low-cost camera modules that is planned for the overall automation system.

This work also attempts to simulate image captured on a purpose-built sorting conveyor belt, where the moving belt is white in colour to assist in the contrast needed for the image capturing. Hence the images captured for this project have a white (but generally dirty or slightly off white) background.

### *3.1 Dataset*

There were initially more than 900 images in JPEG format provided by Melangking Oil Palm Plantation. Since the images were captured by the worker who manually classified the palm fruit bunches, many of the images were discarded due to the images being unusable. The images were discarded due to various reasons such as partial fruit bunch captured or the background was too noisy (littered with loose fruits or leaves). In the planned automated system, the fruit bunches will be on the conveyor belt and there will not have partial image capture as the camera will capture the whole conveyor belt area and there will be minimal loose fruits (fruitlets) as it is planned that the conveyor belt will have a mechanism to flush out the fruitlets. The final dataset[1] used consists of 514 images classified into 6 classes:

[1]Dataset is available at https://www.teradatauniversitynetwork.com/Library/Items/Datasets-from-Melangking-Palm-Oil-Corporation.

- Empty Bunches (57 images)
- Ripe Bunches (190 images)
- Dirty Ripe Bunches (80 images)
- Rotten Bunches (62 images)
- Under Ripe Bunches (53 images)
- Unripe Bunches (72 images).

Although the images were classified by experienced sorting workers, some human errors were expected. The classifications were then validated by the palm fruit bunch sorting supervisor to ensure correctness. Since the image capturing was not controlled, preliminary processing will be required as the background would be dirty and there will be lighting differences due the fact that the images were captured at varying times of the day.

## 4 Methodology

The project employs the following process (Fig. 1).

1. Data pre-processing
2. Feature extraction
3. Modeling.

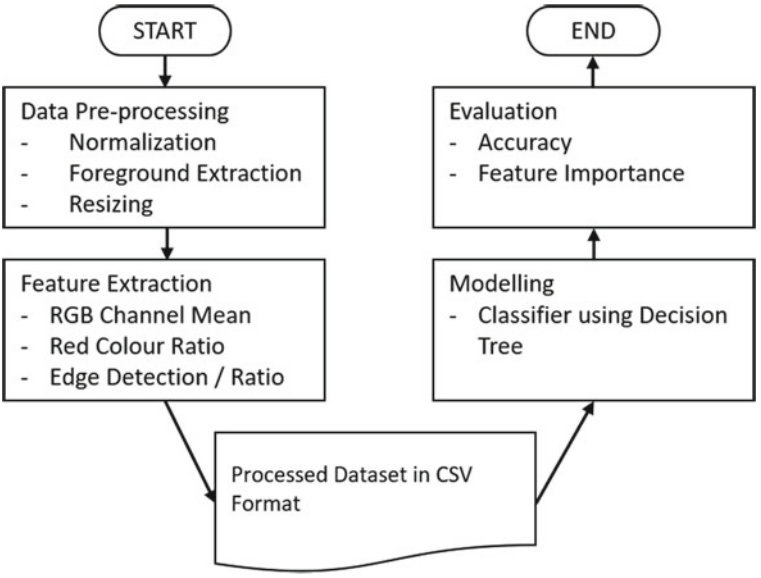



**Fig. 1** Methodology

### *4.1 Data Pre-processing*

The data pre-processing consists of normalization and foreground extraction.

**Normalization.** Due to the varying lighting conditions for each of the images captured, a normalization process is required. The uncontrolled environment caused varying saturation of the images. The Hue Saturation Intensity (HSI) values of the images were used for the normalization. This is based on the centre pixel intensity values and the resulting image will have reduced glare to enable a fairer comparison of the images.

**Foreground Extraction.** The next process is to remove the background of the images. The backgrounds of the images are not useful in determining the classification process. The images were captured in an actual production environment and hence the white background used is smeared with dirt and some may even have loose fruits.

The technique used to remove the background is the GrabCut foreground extraction method developed by Rother, Kolmogorov, and Blake [9]. GrabCut, available in the OpenCV library (https://opencv.org/), is a segmentation algorithm that utilizes edges and region detection in order to extract the foreground wanted. GrabCut also uses a system where a user may mark a certain region as foreground or background. This can be done using by either manually marking the image using a mask, or setting a rectangle in order to capture the foreground region. The latter was chosen as the method was more dynamic and suitable for the large number of images needed to be processed.

Figure 2 shows the result of the palm fruit bunch extracted from the background. The unsupervised GrabCut method is able to pre-process about 90% of the images

**Fig. 2** Example image with background removed

successfully. For the images that were not processed correctly, some pre-processing to define the rectangle for the GrabCut process were done. The images were then resized automatically to 900 pixels wide. This is in order for the modeling to execute within a reasonable time frame.

### *4.2 Feature Extraction*

There are five features that are used, namely the mean of each of the Red, Green and Blue colour channels, the ratio of the red colour in the image and the amount of edges detected.

The images were initially converted to HSV (Hue, Saturation, Value). This is to simplify the manipulation and to enable the usage of a mask. A mask is then created for each image to utilize a threshold to remove pixels. The threshold ranges of Hue values used for the colour of red is 0-10 for the lower range and is set to 180 for the upper range.

**RGB Colour Mean.** The most reliable feature indicating the ripeness of a palm fruit bunch is the amount of the colour "light red". The other colours would be influential in determining the other classifications (other than ripe) and hence the work processed the means of each of the colour channels. A mask was applied to the image to ensure that the result of the mean function would be an accurate average of each colour channel while excluding the black pixels (the mask) of the background.

**Red Colour Ratio.** In order to determine the ripeness, the ratio of the amount of red in the image was determined. The approach used here is to find the ratio of the red pixels in the image and divide it with the total number of pixels of fruit bunch (the foreground that was extracted). This ratio will then be used as a feature for the classification model.

Figure 3 depicts the normalized image where the red pixels are to be extracted from, whilst Fig. 4 depicts the image after applying the threshold values to extract the red pixels. From visual checks, the process was able to successfully extract the palm fruit kernels. Figure 4 is illustrated to also show that even for un-ripe palm fruit bunches that consists of dark kernels, the thresholds applied was sufficient to differentiate the un-ripe fruit kernels. The thresholds are not only for the colour red but also for the cut-off value for the lower end intensity.

**Edges (Spikiness).** The spikiness feature that is being considered in this work is defined as the magnitude of the edges belonging to the palm fruit bunches. This feature was considered due to there being a visually observable difference in roughness of the silhouette of an empty jagged palm fruit bunch and a ripe palm fruit bunch. Thus, canny edge [10] detection algorithm in the OpenCV library [11] was applied to the images (Fig. 5).

To determine the spikiness, the work counted the number of pixels that are considered as edges. The rationale is that the longer the length of the spikes, there will be

**Fig. 3** Normalized image prior to extracting red pixels

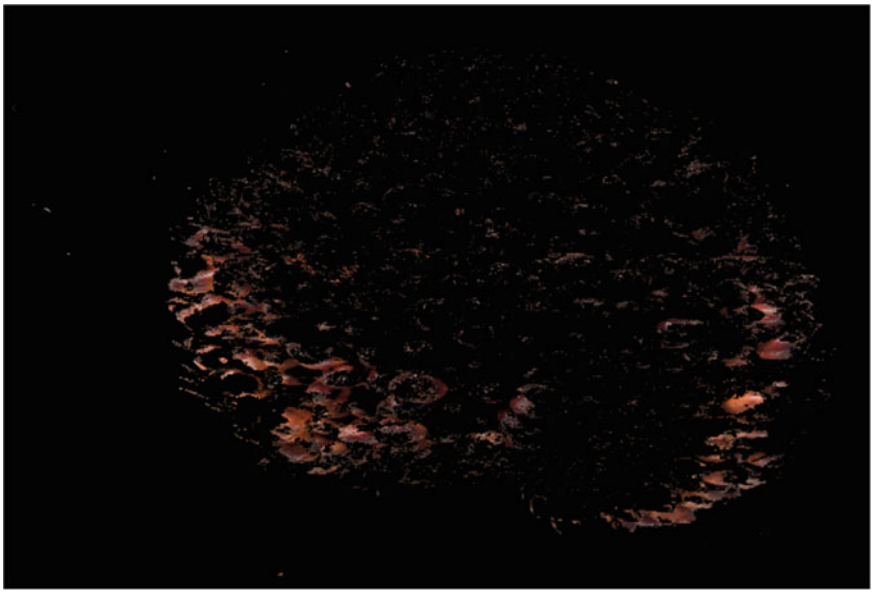


**Fig. 4** Extracted red pixels from normalized image

**Fig. 5** Edge detection of a palm fruit bunch

more pixels used to denote the edges. The number of pixels counted is divided by the total number of pixels in the fruit to obtain the ratio.

## *4.3 Modeling*

After refining the data acquisition methods, the acquired data was stored in a comma-separated values (CSV) file. The data in the CSV files were then normalized to be in the range of 0-1 based on the minimum and maximum value of each attribute (Min-Max Normalization). Instead of using 6 different classes, the images were evaluated based on ripe and unripe as well as ripe, under-ripe and unripe.

For this work, we decided on the Decision Tree (DT) method as the processed dataset is not large and the DT method reflects how a manual plantation worker decision is made, that is from the colour visualized and the spikiness of the palm fruit bunch.

The data was then split into training and test sets where 70% of the data is allocated for the training set and 30% of the data for the test set.

## 5 Evaluation: Results and Discussion

The experiment was conducted 15 times without the setting of the random seed. Results produced by the data modeling of the processed CSV file gave an average accuracy score of 71.11% (Table 1). The accuracy score was computed from the sum of the correctly classified samples over the total sample population.

A noteworthy observation is that the classifications of ripe, under-ripe, and unripe resulted in a lower average accuracy score compared to the performance of binary classification of ripe and unripe. It stands to reason that it would be lower, as it gets more demanding on the decision tree algorithm with the addition of another class.

Observing the confusion matrix (Fig. 6) indicated that ripe fruits have the best performance with about 87% accuracy rating (62 of the 81 samples classified correctly). However, the unripe and under-ripe do not perform as well, with about 50% of the test images were confused to be under the ripe category (13 divide by 28 samples classified correctly). This could be caused by insufficient features for the algorithm to correctly split the set. Moreover, there are quite a few images that are questionable as to whether the images were incorrectly labeled, in other words data noise caused by human error.

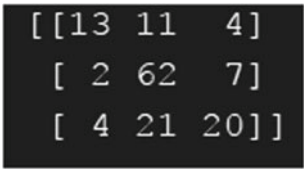
The low accuracy in classifying ripe/unripe and ripe/under-ripe could be caused partially by the number of features being insufficient data for the model to correctly classify between under-ripe and unripe. Moreover, after some observation, it was observed that the colour scheme is somewhat similar when comparing under-ripe fruits and unripe fruits.

Our current method of determining spikiness using edge detection has a moderately significant influence as empty and rotten fruit bunches tended to be spikier. However, our current definition of spikiness can be further improved.

**Table 1** Classification accuracy

| | Maximum (%) | Average (%) |
|---|---|---|
| ripe/unripe | 79.16 | 71.11 |
| ripe/under-ripe/unripe | 68.75 | 64.67 |

**Fig. 6** Confusion matrix for classifier



## 6 Conclusion

The work presented here provides two main contributions. Firstly, the images can be obtained without specialized cameras or in a controlled environment. The use of colours have been well documented and the work here, used the means of the RGB channels independently and the ratio of the red colour itself. Secondly, although the spikiness could do with a better method, using the Canny Edge detection algorithm and conducting the feature engineering of the spikiness factor using the edge to image ratio has shown a significant contribution to the classifier. With a binary classifier, the work presented here is able to achieve up to 79.16% accuracy and with a tri-class classifier, it was able to achieve up to 68.75% accuracy.

## References

1. Meyer G, Mehta T, Kocher M, Mortensen D, Samal A (1998) Textural imaging and discriminant analysis for distinguishing weeds for spot spraying. Trans ASAE 41(4):1189
2. Abbaszadeh R, Rajabipour A, Sadrnia H, Mahjoob MJ, Delshad M, Ahmadi H (2014) Application of modal analysis to the watermelon through finite element modeling for use in ripeness assessment. J Food Eng 127:80–84
3. Choong TS, Abbas S, Shari AR, Halim R, Ismail MHS, Yunus R, Ali S, Ahmadun FR (2006) Digital image processing of palm oil fruits. Int J Food Eng 2(2)
4. Ghazali KH, Samad R, Arshad NW, Karim RA et al (2009) Image processing analysis of oil palm fruits for automatic grading. In: Proceedings of the international conference on instrumentation, control & automation
5. Shabdin MK, Shari ARM, Johari MNA, Saat NK, Abbas Z (2016) A study on the oil palm fresh fruit bunch (FFB) ripeness detection by using hue, saturation and intensity (HSI) approach. In: IOP conference series: earth and environmental science, vol 37, p 012039. IOP Publishing
6. Saaed OMB, Alfatni MSA, Shariff ARM, Hawedi HS (2019) Modeling ripeness grading of palm oil fresh fruit bunches through image processing using artificial neural network. Geoscience Publications. https://www.geosp.net/?p=6896
7. Jaffar A, Jaafar R, Jamil N, Low CY, Abdullah B et al (2009) Photogrammetric grading of oil palm fresh fruit bunches. Int J Mech Mechatron Eng 9(10):7–13
8. Fadilah N, Mohamad-Saleh J, Abdul Halim Z, Ibrahim H, Ali S, Salim S (2012) Intelligent color vision system for ripeness classification of oil palm fresh fruit bunch. Sensors 12(10):14179–14195
9. Rother C, Kolmogorov V, Blake A (2004) "grabcut" interactive foreground extraction using iterated graph cuts. ACM Trans Graph (TOG) 23(3):309–314
10. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 6:679–698
11. Gregori E (2012) Introduction to computer vision using opencv. Embedded Vision Alliance

# Attention Models for Sentiment Analysis Using Objectivity and Subjectivity Word Vectors


**Wing Shum Lee, Hu Ng, Timothy Tzen Vun Yap, Chiung Ching Ho, Vik Tor Goh, and Hau Lee Tong**

**Abstract** In this research, we look at the notions of objectivity and subjectivity and create word embeddings from them for the purpose of sentiment analysis. We created word vectors from two datasets, the Wikipedia English Dataset for objectivity and the Amazon Product Reviews Data dataset for subjectivity. A model incorporating an Attention Mechanism was proposed. The proposed Attention model was compared to Logistic Regression, Linear Support Vector Classification models, and the former was able to achieve the highest accuracy with large enough data through augmentation. In the case of objectivity and subjectivity, models trained with the objectivity word embeddings performed worse than their counterpart. However, when compared to the BERT model, a model also with Attention Mechanism but has its own word embedding technique, the BERT model achieved higher accuracy even though model training was performed with only transfer learning.

**Keywords** Sentiment analysis · Objectivity · Subjectivity · Word vectors

W. S. Lee · H. Ng (✉) · T. T. V. Yap · H. L. Tong
Faculty of Computing & Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
e-mail: nghu@mmu.edu.my

W. S. Lee
e-mail: leews.sam@gmail.com

T. T. V. Yap
e-mail: timothy@mmu.edu.my

H. L. Tong
e-mail: hltong@mmu.edu.my

C. C. Ho
Department of Computing and Information Systems, Sunway University, 47500 Petaling Jaya, Malaysia
e-mail: peterh@sunway.my

V. T. Goh
Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia
e-mail: vtgoh@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_5

## 1 Introduction

The need for language processing is rapidly growing and its use is on the rise. Google, a web search engine, uses language processing to process a large scale of queries every second. As an example, when one starts typing in "Natural Language", Google would suggest the word "Processing" which is the work of language processing. The language processing task is a very immense field, one its sub-domain is sentiment analysis, which determines the opinion or feeling, and polarity of a given piece of text.

In language, there are only two types of statement, a fact, and non-fact statement. Where facts are unchangeable even when the point of view or stand is different, there are called objective statements. On the other hand, the subjective statement is non-neutral and non-fact, which is then very based on comments and feeling at a particular moment in time, which may change over time. This research work looks into how much the sentiment (polarity) of a sentence will be affected models trained with either objectivity and subjectivity datasets are tested with their opposites. A model incorporating Attention Mechanism by Vaswani et al. [1] is proposed for sentiment analysis with models trained with subjectivity and objectivity work vectors is proposed and investigated.

## 2 Literature Review

Collobert claimed that proper training and good quality of word vectors is a boost on accuracy for natural language processing (NLP) tasks (which includes sentiment analysis) [2]. Word embedding can be categorized into two major categories, **context- independent** and **context-dependent** embedding. The differences between context-independent and context-dependent embedding is that context-independent embedding does not take into account the consequences of the ordering of the words in a sentence.

For context-independent embedding, there is Word2Vec by Mikolov et al., [3], improvements and the idea of Neutral Net Language Model (NNLM) by Bengio et al. [4] and Collobert et al. [5], as well as FastText by Bojanowski et al. [6], which is also an improvement on Word2Vec through the use of n-grams and was able to perform better in Word Similarity tasks on a majority of languages and had shown huge improvements on morphology rich languages such as German, in particular datasets GUR350 and GUR65 by Gurevych [7], and ZG222 by Zesch and Gurevych [8]. However on morphology poor language such as English, in particular dataset WS353 by Finkelstein et al. [9], there was significant improvement.

For context-independent embedding, there are techniques such as Deep Contextualized Word Representations or Embeddings from Language Models (ELMo) by Peters et al. [10], which is built on top of Long Short Term Memory (LSTM) neural nets. The ELMo model was able to show significant results in outperforming previous

state-of-the-art (SOTA0 models on benchmarks such as the Stanford Tree-bank (SST-5) from Socher et al. [11]. There is also Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. [12], built on top of Transformers and Attention Mechanism [1]. BERT is not an embedding but a language model and it was able to outperform ELMo on General Language Understanding Evaluation tasks (GLUe) in the work of Wang et al. [13]. There is also RoBERTa: A Robustly Optimized BERT Pretraining Approach by Liu et al. [14] which is an improved version of BERT. The differences in approach is that when the training size tends to get bigger, and BERT is equipped with the Next Sentence Prediction (NSP) and Masked Language Modelling (MLM) tasks, RoBERTa omits the NSP task and utilises a dynamic masking pattern instead of static. RoBERTa was able to improve the accuracy of BERT by few percent.

Although various word embeddings have been around and investigated in various capacities, this research work looks into context-independent embedding on subjective and objective statements, particularly Word2Vec which has been the baseline for most context-independent techniques. It has also been proven to be fast and able to provide high accuracy. In this aspect, Tkachenko et al. [15] proposed SentiVec which looked into improving accuracy from embedding refinement in contrast to this research work which focuses on model implementation.

## 3 Methodology

### *3.1 Datasets*

Two large enough corpora are selected to represent objectivity and subjectivity datasets respectively. The Wikipedia English Dataset [16] is selected as the objectivity dataset while the Amazon Product Reviews Data by McAuley and He [17] is selected as the subjectivity dataset.

The Wikipedia English Dataset is selected because Wikipedia requires all the articles to be factual and have neutral points of view (NPOV), in addition to being large enough to train the models. Amazon Product Reviews is selected because the entries are user experiences, which makes them potentially judgemental and opinionated.

For this research, we assume that all comments made are purely factual or non-factual. For testing, we have selected the Rotten Tomatoes Movie Reviews by Pang et al. [18], an objectivity dataset, as the reviews were editor-picked and movies were reviewed with strict guidelines and must be of NOPV. Subjectivity testing will be the subject of a future investigation.

### 3.2 *Data Preparation*

Pre-processing includes punctuation removal, stop word removal, lemmatization, case lowering and normalization (normalizing non-English text into English text). Two architectures are available for Word2Vec, Skip-gram, and Continuous Bag-of-Words [3]. The former is utilized as it has has shown better results in a comparison with the latter. For optimization, Negative Sampling (NS) is selected over Hierarchical Soft-Max as the former is more widely used [3].

For both datasets (objectivity and subjectivity), they are trained into embeddings of 300d (300 dimension), parameter of 5 negative samples, window size of 5 tokens, removal of short sentences (less than 5 tokens) as-well-as rare words (less than 10 occurrences). Both embeddings are trained for 10 iterations (10 rounds of fitting and tuning).

For each category in the Amazon Product Reviews dataset, the reviews are split into 5000 positive and negative reviews respectively for the sentiment analysis. Each category is also allotted 10,000 unsorted reviews for subjectivity analysis. There are a total of 24 categories. The Rotten Tomatoes Movies Review dataset is split into 5331 positive reviews and 5331 negative reviews, for a total of 10,662 reviews for sentiment analysis. Figure 1 shows the embedding process of the datasets.

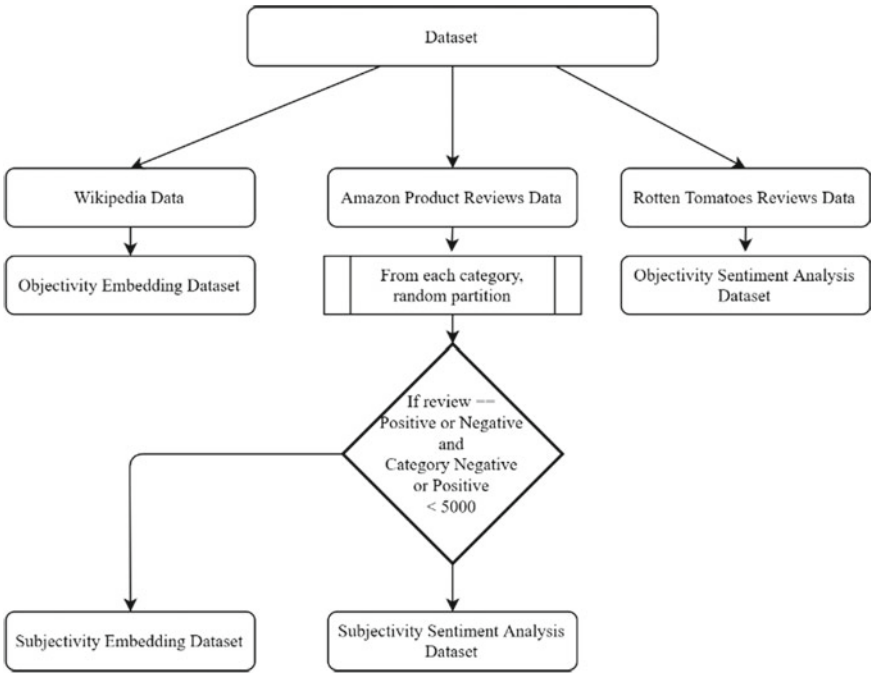



**Fig. 1** Train and test set splits

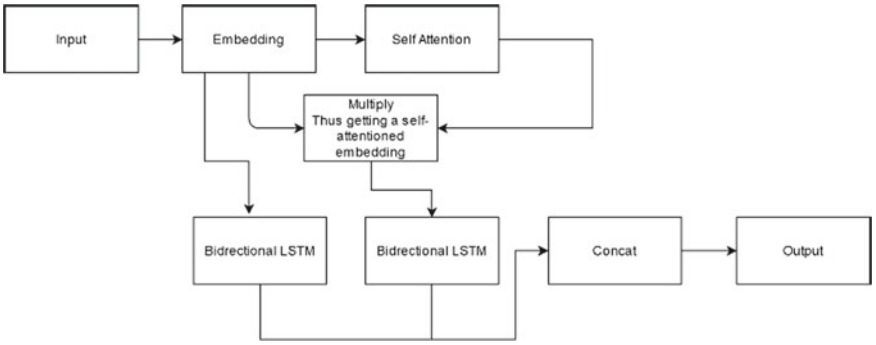



**Fig. 2** Architecture of the proposed model

## *3.3 Model Selection*

To avoid over-fitting or one model favoring one embedding, multiple models are considered, in this case, Logistic Regression and Linear Support Vector Classification (SVC). Previous comparison on subjectivity and objectivity word vectors on these two models have been performed by Tkachenko et al. with the SentiVec embeddings [15]. Normally, a sentence vector is created after the creation of word vectors [19]. However, we believe that some characters may not impose any weight or have any effect, thus an Attention Layer [1] is created instead. Self-attention is a mechanism that is able to assign 'attention' to a key vector (important word). This allows the architecture in a sense to emphasis attention-ed vectors [20].

As such, a model incorporating an Attention Mechanism (Attention Model) is proposed, and its architecture is shown in Fig. 2. The word vectors go through the Attention Layer, generating attention-weighted features. With Bidirectional LSTM neural nets, both the original embedding and the attention-weighted embedding are concatenated to produce sentiment features. The model will be compared against Logistic Regression and Linear SVC.

## *3.4 Design of Experiments*

The three models (Logistic Regression, Linear SVC and Attention) are trained with the objectivity (Wikipedia) and subjectivity (Amazon) embeddings. For training, 10-fold cross-validation is utilized. The models are then tested against the objectivity (Rotten Tomatoes) test set to eliminate the bias of overfitting.

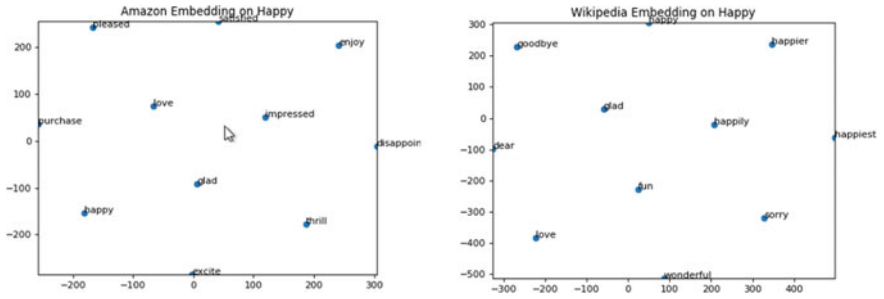



**Fig. 3** t-SNE plots of 'happy' on the Amazon Product Data embedding (left) and Wikipedia English Dataset embedding (right)

## 4 Results and Discussions

### *4.1 Quality of Embeddings*

Figure 3 shows the t-distributed Stochastic Neighbor Embedding (t-SNE) plot for Amazon Product and the Wikipedia English Dataset embeddings on the top 10 nearest words to the word 'happy'. The t-SNE for both datasets show that word similarities are found in the embeddings, for example, 'love', 'fun', 'glad' are clustered together with 'happy'. However, outliers are also present. For Amazon Product Reviews, we can see that 'disappointment' is included, but not found in the t-SNE for Wikipedia English Dataset. On the other hand, For Wikipedia English Dataset, the word 'happy' comes with 'sorry', which is not found for in the t-SNE for Amazon Product Reviews.

### *4.2 Sentiment Analysis*

Table 1 shows the accuracy of the three models trained with the objectivity and subjectivity embeddings. Performance of all three models were very close, with Linear SVC performing the best for objectivity embedding, while Logistic Regression and Attention tied for subjectivity embedding. An interesting observation is that the models with objectivity embedding performed worse than those with subjectivity embedding, even though the testing was performed with an objectivity embedding test set.

**Table 1** Accuracy of the Logistic Regression, Linear SVC and Attention models

| Model | Logistic regression | Linear SVC | Attention |
|---|---|---|---|
| Objective embedding (Wikipedia) | 0.7262 | 0.7286 | 0.7189 |
| Subjective embedding (Amazon) | 0.7637 | 0.7618 | 0.7637 |

**Table 2** Accuracy of the Logistic Regression, Linear SVC and Attention models with data augmentation, and BERT model

| Model | Logistic regression | Linear SVC | Attention | BERT |
|---|---|---|---|---|
| Objective embedding (Wikipedia) | 0. 7203 | 0.7293 | 0.7306 | 0.7805 |
| Subjective embedding (Amazon) | 0. 7622 | 0.7608 | 0.7772 | |

From the performance, we believe there is a limiting factor in the form of the size of the training data. In view of this, we have decided to augment the data. In order to to do this, the corpus was augmented with the Easy Data Augmentation (EDA) [21]. Augmentations include synonym replacement, random insertion, random swap, and random deletion. The performance of the models after augmentation are shown in Table 2. The Attention model achieved the highest accuracy, although the difference between the other two models were not big.

For comparison, we also considered BERT [12]. In BERT, transformers are equipped with attention mechanism, encoder and decoder and has achieved high accuracy on the General Language Understanding Evaluation (GLUE) benchmark by Wang et al. [13]. In the case of BERT, the uncased English language model (BERT's default trained with Wikipedia and BookCorpus by Zhu et al. [22]) was utilized and transfer learning was applied, thus none of the previous embeddings were used. In transfer learning, instead of re-creating the whole network architecture, it uses the base architecture and fine-tune the last few layers. This allows for acceptable performance in shorter time because instead of training the whole new network architecture, only a small number of layers on the network architecture is trained. Transfer learning has shown improvement and great accuracy on multiple areas, not limited to text, such as safety guardrail detection that uses the base architecture from VGG-16 network by Kolar et al. [23], melanoma screening using the base architecture from ImageNet by Menegola et al. [24] and for text, there is BioBERT, a BERT architecture that is trained for Medical Usage by Lee et al. [25].

With the usage of transfer learning, there is a reduction in computation and training time. The result from BERT is appended to Table 2 for comparison with the other results. BERT has the best performance compared to the other three models. Even though transfer learning was applied on the BERT model, it still performed the best out of all the four models. We believe that this is the case because of the large pre-trained data for its embedding, and its encoder and decoder structure. While the Attention Model is only encoded with one attention-mechanism, BERT utilizes the attention-mechanism in multiple steps.

## 5 Conclusions

Word embeddings representing the notion of objectivity and subjectivity were utilized in models for sentiment analysis. The models utilized include Logistic Regression,

Linear SVC and the proposed Attention model. The Attention model was able to achieve higher accuracy compared to the other two models when a large enough dataset was used. In the case of objectivity and subjectivity, models trained with the objectivity word embeddings performed worse than their counterpart. When compared with a state-of-the-art model, BERT was able to achieve higher accuracy compared to the other three models even though the model training was performed with only transfer learning.

In future, we will look into testing with a subjectivity dataset on both word embeddings. Furthermore, incorporating the notion of objectivity and subjectivity into the embeddings of the BERT model will also be considered.

**Acknowledgements** This work was supported by the the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme with grant number FRGS/1/2018/ICT02/MMU/03/6 and Multimedia University, under the CAPEX fund with grant number MMUI/CAPEX170008.

## References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
2. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12(Aug):2493–2537
3. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
4. Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3(Feb):1137–1155
5. Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning, pp 160–167
6. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
7. Gurevych I (2005) Using the structure of a conceptual network in computing semantic relatedness. In: International conference on natural language processing, pp 767–778. Springer, Heidelberg
8. Zesch T, Gurevych I (2006) Automatically creating datasets for measures of semantic relatedness. In Proceedings of the workshop on linguistic distances, pp 16–24
9. Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, Wolfman G, Ruppin E (2001) Placing search in context: the concept revisited. In: Proceedings of the 10th international conference on World Wide Web, pp 406–414
10. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv preprint arXiv:1802.05365
11. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1631–1642
12. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

13. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) Glue: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461
14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, …, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
15. Tkachenko M, Chia CC, Lauw H (2018) Searching for the x-factor: exploring corpus subjectivity for word embeddings. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 1212–1221
16. Wikimedia.org (n.d.) Wikimedia Downloads. Wikimedia.org. https://dumps.wikimedia.org/backup-index.html
17. He R, McAuley J (2016) Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th international conference on world wide web, pp 507–517
18. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 conference on Empirical methods in natural language processing, vol 10, pp 79–86. Association for Computational Linguistics
19. Liu H (2017) Sentiment analysis of citations using word2vec. arXiv preprint arXiv:1704.00177
20. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Advances in neural information processing systems, pp 577–585
21. Wei JW, Zou K (2019) Eda: easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196
22. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision, pp 19–27
23. Kolar Z, Chen H, Luo X (2018) Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. Autom Constr 89:58–70
24. Menegola A, Fornaciali M, Pires R, Bittencourt FV, Avila S, Valle E (2017) Knowledge transfer for melanoma screening with deep learning. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), pp 297–300. IEEE
25. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240

# A Question-Answering System that Can Count


**Abbas Saliimi Lokman, Mohamed Ariff Ameedeen, and Ngahzaifa Ab. Ghani**

**Abstract** This paper proposes a conceptual architectural design of Question-Answering (QA) system that can solve "counting" problem. Counting problem is the inability of QA system to produce numerical answer based on retrieved rationale (in text passage) containing list of items. For example, consider "How many items are on sale?" as question and "Currently shampoo, soap and conditioner are on sale" as retrieved rationale from text passage. Normally, system will produce "shampoo, soap and conditioner" as an answer while the ground truth answer is "three". In other words, system is simply unable to perform the counting process needed in order to correctly answer such questions. To solve this problem, QA system architecture with following components is proposed: (1) A classifier to determine if given question requires a counting answer, (2) A classifier to determine if current system's answer is not numeric, and (3) A counting method to produce numerical answer based on given rationale. Despite looking like a whole system, the proposed architecture is actually a modular system whereby each component can operate independently (allowing each component to be separately implemented by other systems). In essence, this paper intents to demonstrate a general idea of how the defined problem can be solved using a modular system, that hopefully also opens up more flexible enhancements in the future.

**Keywords** QA system · Natural language processing · Machine learning

This work was supported in part by Department of Higher Education, Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS), through Universiti Malaysia Pahang (Ref: FRGS/1/2018/ICT02/UMP/02/12).

A. S. Lokman (✉) · M. A. Ameedeen · N. Ab. Ghani
Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, Gambang, Kuantan, Pahang 26300, Malaysia
e-mail: abbas@ump.edu.my
URL: http://www.ump.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_6

## 1 Introduction

Within the software engineering field [16, 17], Question-Answering (QA) system is a computerized algorithm that produces output (answer) based on user's input (question) through the usage of Natural Language Processing (NLP). Relatively similar to chatbot [13–15], QA system will produce answer based on rationale (a span of text) extracted from a given text passage/corpus. To find a correct rationale, system must understand the question's context whether it is direct (no polysemy involved), or indirect (has some sort of polysemy or hidden meaning). To understand context in natural human language, the system must first understand the language itself (primarily the meaning of each word in the sentence). Over the years, NLP researchers have come up with a system called Language Model that can "understand" human language by learning word co-occurance patterns.

Mathematically, Language Model (LM) is a probability distribution system that produces probability values for each word in the text sequence (typically large sequence such as collection of documents). By predicting occurrence probability of words-against-words, relationship of semantic emerges ("a word is characterized by the company it keeps"—[5]). This relationship is what makes LM "understand" human language, that is not perceiving word as an atomic item, but as surrounding words that relate to it. While LM is good at understanding language, it still needs to be fine-tuned in order to perform well in task specific processes (such as QA). This is because LM only understands language but not how to use it (like human understand cars but need to learn how to drive it).

To fine-tune LM for QA system, a specific QA dataset is used. QA dataset generally is a collection of human answers towards human questions in regard to text passages (a reading comprehension exercise). Three most referred QA datasets to date are SQuAD (The Stanford Question Answering Dataset) [19], CoQA (A Conversational Question Answering Challenge) [20] and QuAC (QuAC: Question Answering in Context) [3]. For all datasets, fine-tuned LMs (with additional components) are shown to perform well [8, 9]. In general, evaluation is done based on how well the system performs towards overall data points. With majority correctness as main target, minor error is not given much attention. One of such error is identified as "counting" problem [8].

Counting problem in QA system is an error where the produces answer for "How many" type question is not numerical. As an example, consider the question "How many items are on sale?". With good fine-tune LM, QA system can retrieved following rationale from a given passage "Currently shampoo, soap and conditioner are on sale". Without solution to counting problem, QA system then will answer "Shampoo, soap and conditioner" while in fact the ground truth answer is "Three". Because the QA system is not able to solve the counting problem, it can only produce answer that is a list of relevant items based on identified context in the question.

Following are contributions of this paper:

- A proposal of QA system architecture that can solve the counting problem.
- A proposal of three independent components in modular system setup that can be assembled in order to solve the counting problem.

## 2 Related Work

Current trend in NLP is pretrained Language Model (pLM), that is a generalization of LM. pLM is an LM that can independently be trained to relatively understand any textual human language. After being trained, pLM can be fine-tuned in order to make it perform well on specific downstream Natural Language Understanding (NLU) tasks. Example of such task are single and pair sentence classification, sentence tagging, reading comprehension, and so on [21]. Among currently famous pLM are GPT-3 [2], BERT (and it's variations) [4, 7, 10–12], XLNET [25] and ELMo [18]. For QA systems, "reading comprehension" is the core NLU downstream task that needs to be addressed. To make pLM able to perform well on this task, fine-tuning is done using QA datasets such as SQuAD [19], CoQA [20] or QuAC [3]. In general, those datasets contain three interrelated data: (1) A text passage about a particular subject, (2) A factual question related to the passage (multiple questions per passage), and (3) An answer to each question (to be noted that each dataset has its own quirks but basic data structure is fairly similar).

Fine-tuning pLM might address basic QA system requirement but there is still more to be improved. The basic two steps approach for QA system in answering question is: (1) Looks for factual information (a rationale) in text passage based on semantic information in the question (using learned QA relationship patterns), and (2) Produce answer based on retrieved rationale (rationale is a span of text from the referred passage). Because each dataset/domain has its own quirks and features, researchers augment pLM with various modules to make it perform better in regards to each dataset. Following are some of those augmented implementation: Wen et al. augment pLM with specific module to address why-question in clinical domain QA system [23], Banarjee et al. augment abductive information retrieval method to pLM in order to address open book QA reasoning [1], Wang et al. augment pLM with scoring mechanism for multi-passage answer retrieval in attempt to improve open domain QA system [22], Yang et al. augment specific information retrieval toolkit called Anserini to address QA system with large passage sequence [24], and Godbole et al. augment IR technique that able to address multi-hop QA requirement [6].

Among those augmented modules, none has addressed "the counting problem" issue. Although not presented much in those datasets (only 5.1% in CoQA [20] dataset and even smaller percentage in others), the counting problem is everywhere in everyday human conversation. One such conversation domain is buying-and-selling where "how many" question type/style is used a lot when dealing with items on sale, in stock, in packaging and so on. With this realization, this paper intents to demonstrate

a general idea of how the counting problem can be solved using a modular system. With modularization architecture, proposes system's modules can be augmented into current system as to improve SoTA results on "how many" type questions. Proposes system is also unbiased towards knowledge domain, that is because its targeted on non-contextual word for question type classification (further explanation is in subsection 3.2 "Question type classifier (C1)").

## 3 Architecture

This paper proposes a QA system architecture that is designed upon two main objectives: (1) To solve the counting problem in system's answer, and (2) To become reusable in other systems. While objective one is very clear (the main objective), objective two requires some elaboration. To make a system that is reusable for other systems, it's components need to be modular (independent units). To achieve this, the propose system must maintain the fundamental process of how QA system produce an answer, that is by retrieving rationale from the given passage. By maintaining this process, answers that do not suffer from the counting problem can produce similarly as before. Following Fig. 1 depicts the proposed architecture (the whole system) in data flow format.

Referring to Fig. 1, general input-output process for proposed architecture is fundamentally similar to other QA system where system accepts two input texts (question and passage) and produce one output text (answer). Following the data flow, input will go through four system components: (1) Fine-tune language model (ftLM), (2) Question type classifier (Classifier 1 or C1), (3) Answer type classifier (Classifier 2 or C2), and (4) Counting method (Method 1 or M1). Apart from normal flow (going through all four components), data can also flow directly to the last stage from C1 or C2 component. These are cases where system can bypass subsequent component if it has been identified that the inputs did not suffer from the counting problem (further explanation is in subsection 3.4 "Counting method (M1)"). To be noted that ftLM is a required component for C1 and C2 (required for language and

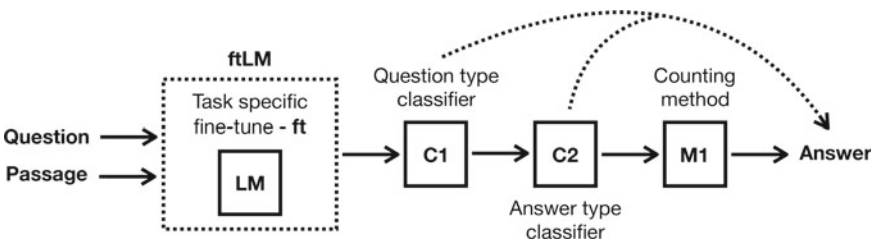



**Fig. 1** General architecture

sentence comprehension task). As such, if C1 or C2 is to be used independently, ftLM is needed to be included as preprocessing module. Next subsection will elaborate more on each component.

### 3.1 *Fine-Tune Language Model (ftLM)*

ftLM is the first component that system's inputs need to go through. ftLM main objective is to retrieve rationale from input passage given input question's context. In details, ftLM needs to perform two interrelated processes which are: (1) Reading comprehension and (2) Rationale retrieval. Referring back to Fig. 1, ftLM (big dashed-line box) contains LM (small solid-line box) surrounded by "Task specific fine-tune" process *ft*. In correlation, LM is reading comprehension (process 1) module, and *ft* is rationale retrieval (process 2) module.

In QA system context, ft is a process of fine-tuning LM using QA dataset such as SQuAD, CoQA and QuAC. Conclusively, LM is used for system to comprehend pre-trained language (e.g. English language) and ft is used for system to guess semantic relation between question text and passage text. As overall architecture is designed to be modular, rationale extraction can be improved by embedding more component towards base ftLM module.

### 3.2 *Question Type Classifier (C1)*

C1 is a component that takes question and passage text as inputs, and classify it into one of two question types: (1) Question that does not requires numeric answer, and (2) Question that requires numeric answer (a binary classification problem). Following Fig. 2 depicted the propose C1 component design.

Figure 2 denoted ftLM as fine-tune Language Model (as in Fig. 1), Q as question text, R as ftLM's resulting rationale, A1 as preprocessor, A2 as Neural Network (NN), and A3 as Sigmoid function. As explained in the previous section, ftML will retrieve rationale from input passage given input question's context. From ftML, A1 will receive rationale and also the original question text as its inputs. A1 objective is to remove all context-related words from question input in order to get not-in-context

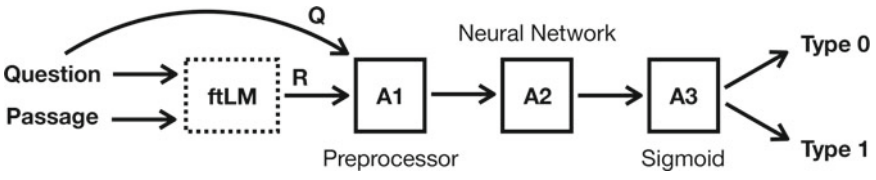



**Fig. 2** Question type classifier

words which is the question words (e.g. "What", "Where", "How many" and so on). Following is the proposed algorithm for A1 subcomponent:

1. Get question text Q and rationale R as inputs
2. Get each word embedding for Q and R
3. Calculate cosine similarity cos($\theta$) between each word in Q, towards each word in R
4. Remove word in Q that cos($\theta$) value is within threshold t
5. Pass remaining Q words embedding to the next process.

Upon completion, A1 subcomponent module will pass embeddings of not-in-context question words into A2. A2 is a multilayer neural network with Sigmoid neuron (A3) at the end. Standard logistic function (where k = 1, L = 1, $x^0 = 0$) will be used for Sigmoid as C1 needs to predict binary class for question type which are: Type 0 - Question that does not requires numeric answer, and Type 1 - Question that requires numeric answer. Sigmoid calculation will produce a value within 0 to 1 range, that is a classification of Type 0 for 0 to 0.49 range value, and classification of Type 1 for 0.5 to 1 range value. For C1 training purpose, QA dataset will needs a label value where Type 0 is labeled 0, and Type 1 is labeled 1 (supervised ML model).

### *3.3 Answer Type Classifier (C2)*

The purpose of C2 component is to classify whether ftLM (that has been fine-tune for QA system) produce a numeric answer or not. Similar to C1, C2 needs to solve binary classification problem with natural language text as its inputs. Following Fig. 3 depicted the propose C2 component design.

Figure 3 denote ftLM as fine-tune Language Model (as in Figs. 1 and 2), B1 as Neural Network (NN), and B2 as Sigmoid function. From ftML, B1 will receive rationale text embeddings as input. Similar to A2 (Fig. 2), B1 is a multilayer neural network with standard logistic function Sigmoid neuron (B2) at the end. Binary classes to be classified by B1 are: Type 0—Answer that is semantically not numeric, and Type 1—Answer that is semantically numeric. To be a "semantically numeric" answer, rationale text must contains numerical-valued text, be it in actual number (e.g. 1, 2, 3) or text representing number (e.g. one, two, three). As for training process, C2

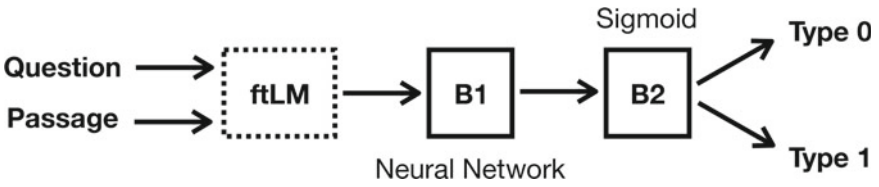



**Fig. 3** Answer type classifier

will use standard QA dataset with additional label value that is similar to C1 training requirement where Type 0 is labeled 0, and Type 1 is labeled 1 (also a supervised ML model).

### *3.4 Counting Method (M1)*

To perform a counting process, C1 and C2 components need to satisfy following two conditions: (1) Input question is a type that requires numeric answer (Type 1 for C1 classifier), and (2) ftLM outputted rationale is not semantically numeric (Type 0 for C2 classifier). By basic logic, condition one is prerequisite to condition two. This is because the counting process is only needed for question that requires a numeric answer. Following Fig. 4 depict path that needs to be satisfied by the data flow in order for M1 to be activated.

M1 component objective is to receive natural language text as input (in a list of items format), and produce the summation of items as output. As powerful as ML is, it is still bound to basic probability principle which is "predicting a value". In mathematical calculation, the produced answer is definite therefore no amount of data can be feed into ML to predicts every possible outcome of mathematical calculation (because real numbers are infinite). As such, M1 requires traditional computing method in calculating summation of items in a list. Following is proposed algorithm for M1 component:

1. Get answer text A as input
2. Segment A text into item array Ar through lexical analysis
3. Loop Ar to count items presented in A (count++)
4. Output final count value.

When complete, M1 will produce a count value (the summation of items) in number format. As QA system usually uses text format, a conversion is needed. In usual formal writing, number 1 to 10 is written in text format while number 11 and above is written in number format. On that account, count value only needed to be converted into text when its value is under 10. Maintaining this scope will make it simpler for conversion process as real numbers are infinite.

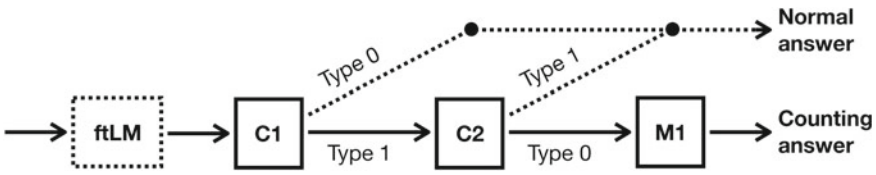



**Fig. 4** M1 activated path

## 4 Discussion and Future Work

This paper proposes a QA system general architecture to solve the counting problem using modular four parts components. This proposed system differs from previous works in two regards: (1) A QA system that specifically intended to solve the counting problem while previous works focused on achieving SoTA result on whole datasets, (2) A modular system component that can also be implemented in other case study (such as binary question/answer text classification), while previous works proposed a close architecture system for specific QA dataset only.

Similar to other conceptual proposal, proof of concept implementation is crucial. For presented architecture, first task is to independently test and verify the three proposed components (C1, C2 and M1). Other than to make sure each component works as intended, this task is also to justify the modular design proposal. After all components are throughly validated, next task is to ensemble all components to become one functional QA system. With this ensemble system, standard QA system benchmark data can be used to test and evaluate the whole architecture. Keep in mind that this system might not achieve SoTA result since its main focus is to solve counting problem. It is however possible for other components to be incorporated later (in order to achieve new SoTA) as the system is modular in design. In addition to standard research datasets, this system can also be validated using real world human conversational datasets as to further justify the needed modules for solving the counting problem in QA domain.

## 5 Conclusion

This paper proposes a conceptual architecture for QA system that intend to solve counting problem in system's generated answer. Proposed architecture is designed to be modular in a sense that each component can work independently. This is to allow other systems to able to embed just the needed component (without the whole system) effortlessly. To solve the counting problem, three components (excluding generalized language model) are proposed within one ensemble system. Those components are (1) Question type classifier, (2) Answer type classifier and (3) Counting method. Component one and two are machine learning based, while component three is traditional programming method. Collectively, all components can be used to solve counting problem while independently, each component can be used to solve its predefined purpose (binary classification for component one and two, and syntactic mathematical addition problem for component three).

## References


1. Banerjee P, Pal KK, Mitra A, Baral C (2019) Careful selection of knowledge to solve open book question answering. arXiv preprint arXiv:1907.10738
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165
3. Choi E, He H, Iyyer M, Yatskar M, Yih WT, Choi Y, Liang P, Zettlemoyer L (2018) Quac: question answering in context. arXiv preprint arXiv:1808.07036
4. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
5. Firth JR (1957) A synopsis of linguistic theory, 1930–1955. Studies in linguistic analysis
6. Godbole A, Kavarthapu D, Das R, Gong Z, Singhal A, Zamani H, Yu M, Gao T, Guo X, Zaheer M, McCallum A (2019) Multi-step entity-centric information retrieval for multi-hop question answering. arXiv preprint arXiv:1909.07598
7. Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) Spanbert: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguistics 8:64–77
8. Ju Y, Zhao F, Chen S, Zheng B, Yang X, Liu Y (2019) Technical report on conversational question answering. arXiv preprint arXiv:1909.10772
9. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942
10. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–40
11. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, Wang P (2020) K-BERT: enabling Language Representation with Knowledge Graph. arXiv preprint arXiv:1909.07606
12. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
13. Lokman AS, Ameedeen MA (2019) Modern chatbot systems: a technical review. In: Arai K, Bhatia R, Kapoor S (eds) Proceedings of the Future Technologies Conference (FTC) 2018. FTC 2018. Advances in Intelligent Systems and Computing. Springer, Cham, vol 881, pp 1012–1023
14. Lokman AS, Ameedeen MA, Ghani NA (2020) A conceptual IR chatbot framework with automated keywords-based vector representation generation. IOP Conf Ser Mater Sci Eng 769(1):012020 IOP Publishing
15. Lokman AS (2011) Chatbot development in data representation for diabetes education. MCS Thesis, Universiti Malaysia Pahang, Pahang, Malaysia
16. Ong MIU, Ameedeen MA, Azmi ZR, Kamarudin IE (2018) Systematic literature review: 5 years trend in the field of software engineering. Adv Sci Lett 24(10):7278–7283
17. Ong MIU, Ameedeen MA, Kamarudin IE (2018) Meta-requirement method towards analyzing completeness of requirements specification. In: Arai K, Bhatia R, Kapoor S (eds) Proceedings of the Future Technologies Conference (FTC) 2018. FTC 2018. Advances in intelligent systems and computing. Springer, Cham, vol 881, pp 444–454
18. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv preprint arXiv:1802.05365
19. Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822
20. Reddy S, Chen D, Manning CD (2019) CoQA: a conversational question answering challenge. Trans Assoc Comput Linguistics 7:249–266
21. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461
22. Wang Z, Ng P, Ma X, Nallapati R, Xiang B (2019) Multi-passage bert: a globally normalized bert model for open-domain question answering. arXiv preprint arXiv:1908.08167

23. Wen A, Elwazir MY, Moon S, Fan J (2019) Adapting and evaluating a deep learning language model for clinical why-question answering. arXiv preprint arXiv:1911.05604
24. Yang W, Xie Y, Lin A, Li X, Tan L, Xiong K, Li M, Lin J (2019) End-to-end open-domain question answering with bertserini. arXiv preprint arXiv:1902.01718
25. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237

# Contactless Patient Authentication for Registration Using Face Recognition Technology


**Kian Yang Tay, Ying Han Pang, Shih Yin Ooi, and Fan Ling Goh**

**Abstract** Patient registration is an essential process in every clinic and hospital before services are provided to patients. Usually, patient's identity card or fingerprint (through a fingerprint scanner) will be requested for identity authentication in order to retrieve medical records of the patient. However, the current global health crisis of COVID-19 pandemic is raising concern on the hygiene and safety of sharing objects or touching surfaces. Same worry is also occurred towards the patient registration interaction process; further, hospitals and clinics are classified as high risk premises. Therefore, a contactless patient authentication for registration using face recognition technology is proposed in this work. In this system, a face is scanned and processed. If the face exists in the database indicating that the subject is an established patient, the patient's records will be retrieved. Else, a new patient registration will be performed to register a new account. The efficiency of the system is assessed using our self-collected database. Empirical results show that the proposed system is able to attain 94% accuracy. But, an inferior performance is obtained, especially dealing with makeup variation.

**Keywords** Contactless · Patient registration · Face recognition · Local binary pattern

K. Y. Tay · Y. H. Pang (✉) · S. Y. Ooi
Faculty of Information Science and Technology, Multimedia University, Ayer Keroh, 75450 Melaka, Malaysia
e-mail: yhpang@mmu.edu.my

K. Y. Tay
e-mail: 1151105440@student.mmu.edu.my

S. Y. Ooi
e-mail: syooi@mmu.edu.my

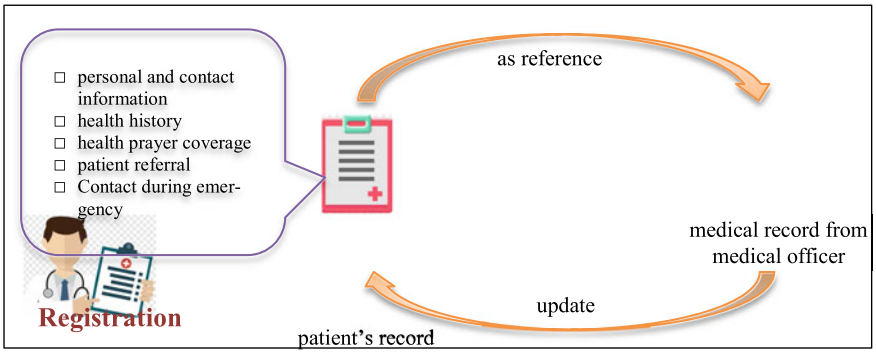
F. L. Goh
DrSoft Sdn Bhd, Taman Malim Jaya, 75250 Melaka, Malaysia
e-mail: loverpet.stella@gmail.com



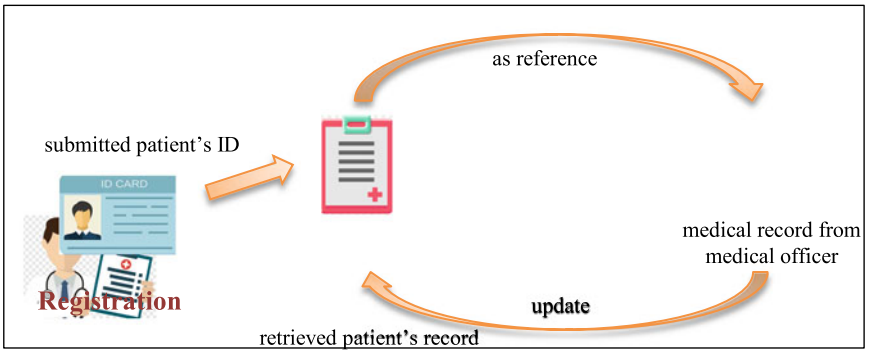
R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_7

## 1 Introduction

Patient registration is essential to the bottom line of a medical practice. It is a complex process that collects a substantial amount of patient's data, including demographic information, health payer coverage, health history, etc. Patient registration is the first step to generate medical record and personal information of the patient before healthcare services could be provided. It is a mandatory process in every clinic and hospital in order to provide services to the patient, as well as to keep a record of the services that has been availed by the patient. For an established patient, the patient is also required to perform registration at the clinic counter as the first step of each visit (revisit). Usually, it is done by just submitting his identity credential for identity authentication. Once the patient's identity is verified, the registration is to confirm and update the patient's demographics and insurance information due to the frequently changes of circumstances, as well as retrieving the medical record of the patient which is crucial as a reference for medical officer's diagnosis. Figure 1 illustrates




(a)




(b)

**Fig. 1** Registration process of (**a**) new patients and (**b**) established patients

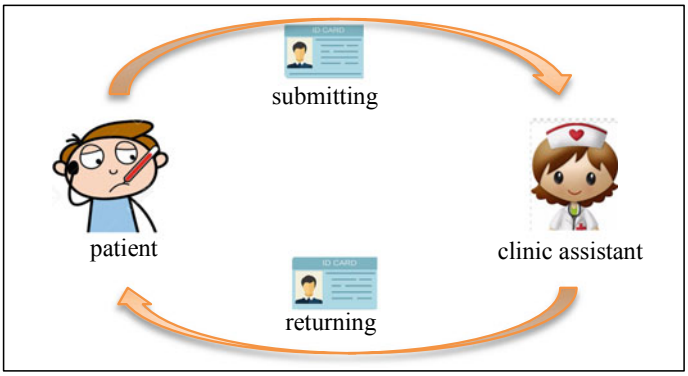



**Fig. 2** Registration interaction between a patient and a clinic assistant

new and returning patient registration processes.

A patient is required to submit his identity card, coined as IC, to a clinic registration assistant for identity authentication. With holding the submitted IC, the clinic assistant will help the new patient for registration or retrieve the medical record of the established patient. This registration process has been practiced for years. Figure 2 illustrates the registration interaction between a patient and a clinic assistant.

In some healthcare institutes, on top of handling the IC, patients are required to scan their fingerprints via a fingerprint scanner for multimodal identity authentication. However, the recent global health crisis of COVID-19 pandemic, also known as coronavirus pandemic, is raising concern on the hygiene and safety of sharing objects or touching surfaces in public spaces. Same worry is also occurred towards the patient registration interaction process; further, hospitals and clinics are classified as high risk premises.

COVID-19 virus is transmitted between people through respiratory droplets and contact routes [1–5]. The *virus* remains suspended in *droplets* smaller than 5 micrometers, known as aerosols. It can stay suspended and live on surfaces from several hours to a few days [6–8]. These touching surfaces are paper, copper, cardboard, wood, cloth, plastic, glass, etc., which are common touching objects like door knobs, elevator buttons, shopping carts, public faucets, ATM screens/buttons, gas pumps, handrails, etc. Nevertheless, those fingerprint scanners used for patient registration can be likely contaminated. Possibly, the route of virus transmission begins from the carrier's nose, eye or mouth and moves to his fingers, then to the fingerprint scanner. Since fingerprint scanner is a shared device, from there the virus transfers to another person. Handling a card is similar to touching any other surface. If the contaminated card is not sanitized, it can harbor germs and act as a virus transmission medium.

The sign of nervousness over touching surfaces has arisen. In viewing this, the current patient registration system should be revised to minimize the contact procedure. The hygiene distress raises the demand of touchless technology. In this work, a contactless patient registration system using face recognition technology is proposed. In literature, there are numerous deep learning approaches for highly accurate face

recognition [9–12]. With softmax and correlation loss supervision, the proposed deep correlation feature learning (DCFL) model learns deep features with the inter-class separability and the intra-class compactness [9]. This deep model shows highly discriminative capability for face verification. Further, a deep heterogeneous feature fusion network is also proposed [10]. In this network, different deep convolutional neural networks are constructed to generate complementary informative features for template-based face recognition. This approach fuses different deep features by learning the nonlinear projection of the deep features and producing a discriminative representation through preserving the inherent geometry of the deep features. Undoubtedly, the performance of deep learning approaches is great. However, deep learning is highly computational intensive [13]. A high-priced computer with high performance computing is required. This may cause limited potential markets since many panel clinics are hardly to invest such expensive machine just for patient registration purpose. Hence, a local feature descriptor is adopted as feature extraction technique in the face recognition system. Unlike those deep approaches, the implementation of the local feature descriptor is much computational efficient and a common home computer is sufficient to support the process.

## 2 Patient Registration System Using Face Recognition

This patient registration system is developed with the following coding environment:

- Microsoft Windows 10 (64bit)
- 8 GB RAM
- Intel Core i7-7700HQ
- NVIDIA GeForce GTX 1050
- NetBeans IDE
- MySQL database (database setup)
- XAMPP (database setup)
- Java Development Kit (JDK 8)
- OpenCV 2.4 (face recognition)
- JavaCV 1.5 (face recognition).

Figure 3 illustrates the flow diagram of a patient registration process. A face is scanned and processed. If the face exists in the database, this means that the subject is an established (existing) patient of the clinic, so the relevant records of the subject will be retrieved. On the other hand, if the face never exists in the database, a new patient registration will be performed to register a new account for the patient.

In this system, a real-time face detection and recognition is performed using Local Binary Pattern (LBP) on OpenCV and Java. LBP is a simple yet efficient feature descriptor with low computational complexity [13]. LBP encapsulates local constitute of face images by corresponding each pixel with its neighboring pixels. This makes the extracted features to be invariant to illumination variation. Hence, LBP is able to achieve a good performance in face recognition. Figure 4 shows

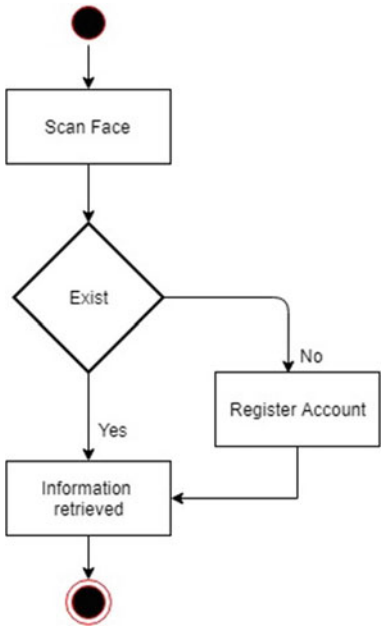



**Fig. 3** Flow diagram of patient registration

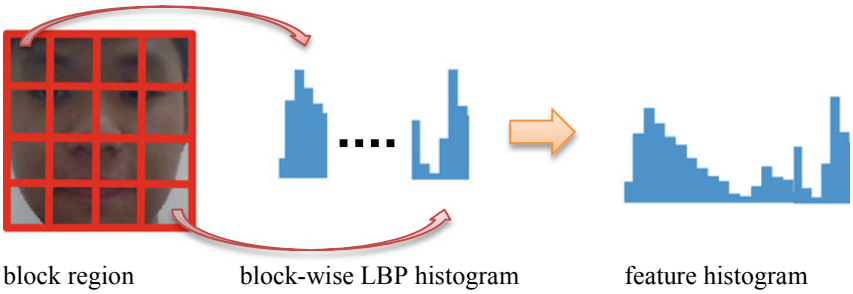



**Fig. 4** Facial feature extraction using LBP

facial feature extraction using LBP. More details of LBP in face recognition could be referred in [14]. Figure 5 illustrates the overview of the face recognition process. A face is detected and captured. The preprocessed face region is further analyzed via LBP descriptor to extract informative features for representation. Then, the extracted feature representation template is classified for identity authentication.

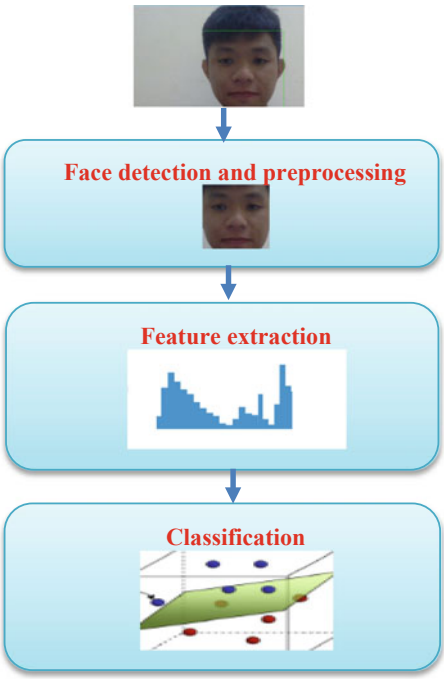



**Fig. 5** The overview of face recognition system

## 3 Application of Patient Registration System

A database is setup to store the data of patients. The data includes patient's personal data, medical records, appointment data as well as their enrolled facial template during the first registration, as shown in Fig. 6. For a returning registration, the newly captured facial data of a patient will be matched with all the facial templates stored in this database. If there is a match, the relevant data of the particular patient will be retrieved within seconds. Else, a new registration page will be prompted out for the new patient to register.

Figure 7 illustrates the pages of the application of patient registration system. Clinic registration assistant has two options to register an existing patient: (1) face

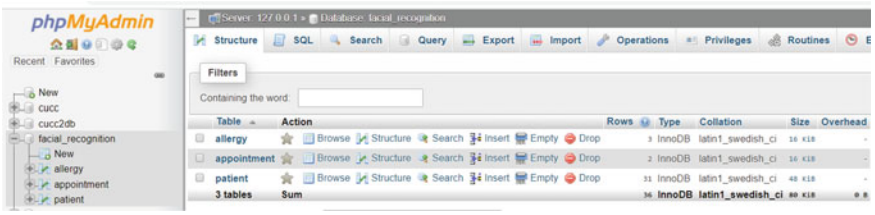



**Fig. 6** Database of patient records

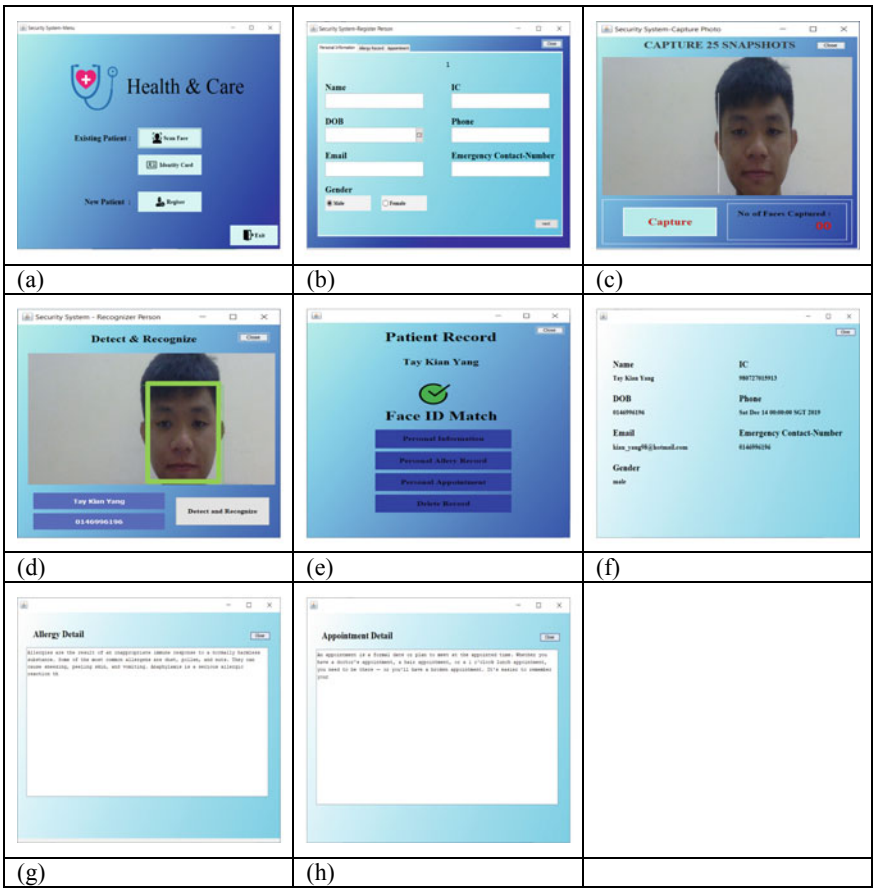



**Fig. 7** Patient registration system

scanning and (2) IC, shown in Fig. 7a. During a new patient registration, personal data as well as medical records of the new patient will be captured and stored in the database, Fig. 7b. 25 facial photos are automatically captured continuously and processed into facial templates which then will be kept in the database for future matching purpose (Fig. 7c). For an existing patient registration, if option face scanning is selected, the system will automatically detect the face of the patient and perform face recognition, Fig. 7e. If the captured face is matched with the facial template in the database, the patient's records will be retrieved within few seconds, Fig. 7f-h. Else, a new patient registration page will be prompted out.

## 4 System Performance Analysis Test and Discussion

In order to evaluate the efficiency of the proposed system, a system performance analysis test is conducted with 30 subjects with 25 photos per subjects for system training purpose. Each subject is required to perform 5 attempts of testing and the average accuracy score is calculated. To better assess the efficacy of the system in dealing with the real world scenario, we introduce several intra-class variations between training and testing environments:

- Illumination—~40% illumination difference
- Makeup—without makeup during training and with makeup during testing
- facial expression—no expression during training and different expressions during testing
- camera-to-subject distance—~30 cm distance from camera during training and ~60 cm distance from camera during testing
- facial details—without and with glasses.

From Fig. 8, we can observe that the proposed system is able to attain 94% accuracy in face recognition. This deduces that the local features encoded in LBP descriptor are well representing the facial data. However, the performance of the system is slightly inferior when dealing with illumination and facial details variations. The performance is degraded for 4% and 7% in the facial details and illumination variation conditions, respectively. Nevertheless, it is still able to obtain 90% and ~87% performance accuracy. On the other hand, when dealing with facial expression and camera-to-subject distance variations, the performance of our proposed system further drops. The accuracy rate is dropped with 17–20%. From the empirical results, we also observe that the system shows worst performance with makeup intra-class variation. The obtained accuracy result is merely 63%. The result reveals that facial makeup could alter the appearance of a person and degrade the performance of an

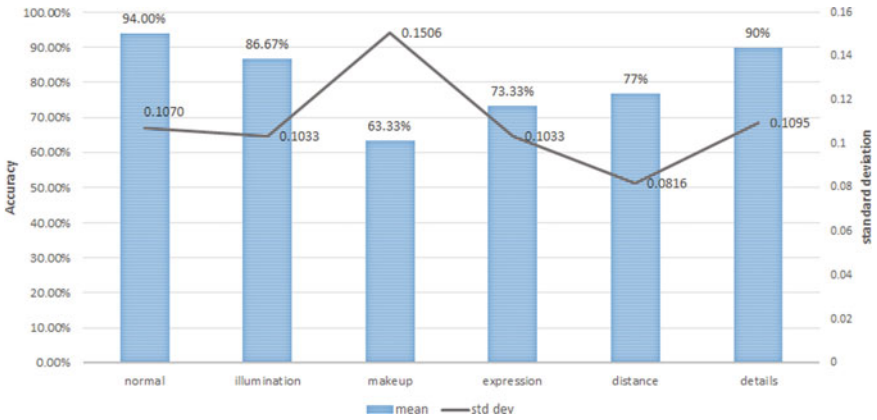



**Fig. 8** System performance accuracy and standard deviation under different conditions

automated face recognition system. This finding is consistent to the literature studies that the performance of their proposed face recognition systems are also affected in the presence of makeup on face [15, 16].

## 5 Conclusion and Future Scope

This paper presents a contactless patient registration system by using face recognition technology. In this registration system, face of a patient is captured and processed into local feature template via Local Binary Pattern descriptor. The transformed feature template will be matched with the facial templates that enrolled earlier (during the new patient registration) in the database. If there is a match, this indicates that the detected patient is an existing patient of the clinic and hence his records will be retrieved. Else, he is required for new patient registration. Experimental results show that the system is able to obtain a promising performance in face recognition when the training and testing conditions are uniform. However, the performance is degraded, especially dealing with makeup variation. In our future work, we will explore a robust algorithm that can well handle intra-class variations, especially makeup variation.

## References

1. Burke RM, Midgley CM, Dratch A, Fenstersheib M, Haupt T, Holshue M, Ghinai I, Jarashow MC, Lo J, McPherson TD, Rudman S, Scott S, Hall AJ, Fry AM, Rolfes MA (2020) Active monitoring of persons exposed to patients with confirmed COVID-19—United States, January-February 2020. MMWR Morb Mortal Wkly Rep 69:245–246
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506
3. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 382:1199–1207
4. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, Xing F, Liu J, Yip CCY, Poon RWS, Tsoi HW, Lo SKF, Chan KH, Poon VKM, Chan WM, Ip JD, Cai JP, Cheng VCC, Chen H, Hui CKM, Yuen KY (2020) A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 395:514–523
5. Liu J, Liao X, Qian S, Yuan J, Wang F, Liu Y, Wang Z, Wang FS, Liu L, Zhang Z (2020) Community transmission of severe acute respiratory syndrome coronavirus 2, Shenzhen, China,. Emerg Infect Dis 26 (2020)
6. Study suggests new coronavirus may remain on surfaces for days—National Institutes of Health (NIH). https://www.nih.gov/news-events/nih-research-matters/study-suggests-new-coronavirus-may-remain-surfaces-days. Accessed 24 May 2020

7. van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A, Williamson BN, Tamin A, Harcourt JL, Thornburg NJ, Gerber SI, Lloyd-Smith JO, de Wit E, Munster VJ (2020) Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. N Eng J Med 382(16):1564–1567
8. Reducing the spread of coronavirus starts with basic hygiene—Harvard Gazette, https://news.harvard.edu/gazette/story/2020/03/preventing-the-spread-of-coronavirus-starts-with-basic-hygiene/. Accessed 24 May 2020
9. Deng W, Chen B, Fang Y, Hu J (2017) Deep correlation feature learning for face verification in the wild. IEEE Sig Process Lett 24:1877–1881
10. Bodla N, Zheng J, Xu H, Chen J-C, Castillo C, Chellappa R (2017) Deep heterogeneous feature fusion for template-based face recognition. In: 2017 IEEE winter conference on applications of computer vision (WACV), pp 586–595
11. AbdAlmageed W, Wua Y, Rawlsa S, Harel S, Hassner T, Masi I, Choi J, Leksut JT, Kim J, Natarajan P, Nevatia R, Medioni G (2016) Face recognition using deep multi-pose representations. In: 2016 IEEE winter conference on applications of computer vision (WACV), pp 1–9
12. Guo G, Zhang N (2019) A survey on deep learning based face recognition. Comput Vis Image Underst 189:102805
13. Chen C, Zhang P, Zhang H, Dai J, Yi Y, Zhang H, Zhang Y, Khan MJ (2020) Deep learning on computational-resource-limited platforms: a survey. Mob Inf Syst (4):1–19
14. Ahonen T, Hadid A, Pietikäinen M (2004) Face recognition with local binary patterns. Lect Notes Comput Sci Lect Notes Artif Intell Lect Notes Bioinform 3021:469–481
15. Dantcheva A, Chen C, Ross A (2012) Can facial cosmetics affect the matching accuracy of face recognition systems? In: 2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS), pp 391–398
16. Kose N, Apvrille L, Dugelay JL (2015) Facial makeup detection technique based on texture and shape analysis. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition, FG 2015. Institute of Electrical and Electronics Engineers Inc

# Drawing and Recognising Simple Shapes with Real-Time Feedback Using Pattern Recognition


**Juharizal Adi Jen, Norizan Mat Diah, and Zaidah Ibrahim**

**Abstract** Pattern recognition is a mature but exciting and fast-developing field concerning computer vision, image processing, shape drawing, and text analysis. The shape pattern can be recognised easily using this technique. Some children have difficulties in identifying shapes. Therefore, through shape drawing exercises, it helps children to understand better. This project aims to develop a mobile application, assisting children in practising drawing using pattern recognition. It can identify shape types by matching the shape pattern with the given input. It will classify the information provided based on the feature extraction using the Freeman Chain Code. Then, each shape pattern is recognised using regular expression tools. Functionality testing has been conducted on this application with an accuracy of 80%. The application will encourage children to draw more by giving feedback on the exercise that they do. It may assist children in learning a new, better way of drawing shape accurately while improving children's fine motor skills.

**Keywords** Children · Drawing · Freeman chain code · Pattern recognition · Shape

## 1 Introduction

Learning about shapes helps children to identify objects as well as letters. For instance, an ice-cream cone is made up of a triangle and one or more circles, while the letter O is made up of a circle. Thus, identifying the object shapes plays an important part in various visual knowledge aspects [1]. Therefore, it is a powerful feature when it can be learnt in the early stage. If children have problems identifying the object

J. A. Jen · N. M. Diah (✉) · Z. Ibrahim
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor Darul Ehsan, Malaysia
e-mail: norizan@fskm.uitm.edu.my

J. A. Jen
e-mail: ardieyadizal@gmail.com

Z. Ibrahim
e-mail: zaidah@fskm.uitm.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_8

shapes, they may have difficulties handling themselves in the real world in the future [2, 3].

It is important to apply both image pre-processing [4] and feature extraction to obtain good results [5]. The pre-processing step involves extracting the x and y coordinates of every pixel while feature extraction involves the Freeman Chain Code creation to represent the shape boundary and edges. The feature extraction ensures every edge point along a contour is being measured so that the shape is perfectly matched. Pattern matching is the final step in which the pattern produced by the Freeman Chain Code is being matched to the collected patterns or shape templates. This template matching technique is important to ensure that the pattern is accurately identified. A regular expression tool is used in this project to match the shape. The regular expression, commonly known as regex, is a special text string for a search pattern. This regular expression has several syntaxes that can be used based on the logical creativity of the programmers.

## 2 Background

A clear basis for problem-solving skills can be achieved if a person can master the pattern. One can identify an object or shape easily through a pattern. The pattern is defined as a design or model used as a guide in needlework and other crafts. It can either be seen physically or observed mathematically by applying algorithms [6]. The more elegant and intricate the pattern, the more beautiful the braid in space-time is, the more satisfying its existence [7].

Informally, pattern recognition is widely used to identify or recognise the related pattern. Pattern recognition is defined as a pattern recognition process by using machine learning or the characteristics detection or data arrangements, providing information about a given system or data set [8].

There are two main categories in pattern recognition, which are supervised and unsupervised learning. In supervised learning, all the data is being labelled, and the algorithms learn the output prediction based on the input data. On the other hand, all data is unlabelled in unsupervised learning, and the algorithms learn to inherent structure from the input data [9].

By using the Freeman Chain Code to extract the shape features, the extraction can be done in every direction. Freeman Chain Code is an external representation used to describe an object or component based on its boundary [10]. If the edge point list along a contour is registered, chain codes are used to represent each object shape borders in the image [11]. The codes sequence is determined by identifying a pixel starting point and the unit vector series. The vectors are obtained from a movement from one pixel to another in several directions; either upward, downward, to the right or left, along the object's boundary until it returns to the starting point. Usually, the Freeman Chain Code has a 4-connected or 8-connected neighborhood (Fig. 1). The chain code contains numbers ranging from 0 to 3 or from 0 to 7 [12].

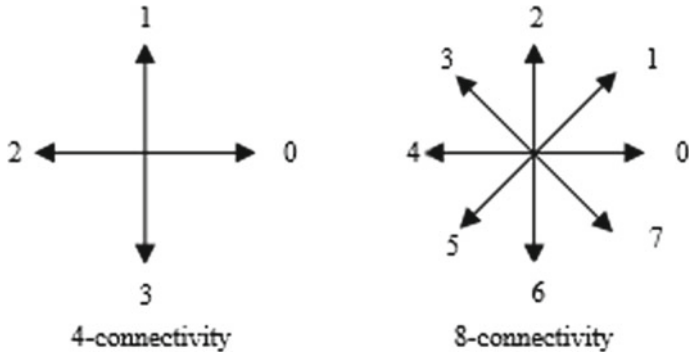



**Fig. 1** The 4-connectivity and 8-connectivity schemes of Freeman Chain Code

Freeman Chain Code has several features that are usable for many researchers. According to [13] Freeman Chain Code can detect circles and arcs. Based on the circle or round objects characteristics in Freeman Chain Code, a full circle will be identified first and followed by the remaining edges.

Azmi and Nasien [14, 15], conducted a study on English handwriting using Freeman Chain Code as data representation, classifying data after preprocessing, and feature extraction stages [16]. Preprocessing involves eliminating noise to facilitate the feature extraction process. The feature extracted as the Freeman Chain Code contains the boundary code of each character image, including the location directly for the next pixel and the corresponding environment in the image. The Freeman Chain Code advantage is that it can reduce data and store information. With this Freeman Chain Code advantage, it has been widely used as a research topic. Furthermore, Freeman Chain Code can detect straight lines. According to [17], the Freeman Chain Code was used in their research because it could recognise lines effectively and accurately. In the study, they used the boundary trace method to get the boundaries in binary images and calculate the Freeman Chain Code for each boundary.

## 3 System Development

The system development begins at the image pre-processing stage starting after gathering the input data from the mobile phone. The input data consists of the shape types. Next, it will pass through the feature extraction stage in which the image gathered are extracted using the Freeman chain code. The obtained data will be inserted into a new data file. The pattern matching stage will follow suit to match the input data from the early stage to see whether it matches. In this stage, the extracted pattern string will be matched to expressions being set up. The results, either successful or unsuccessful, will be displayed on the screen. The system architecture is shown in Fig. 2.

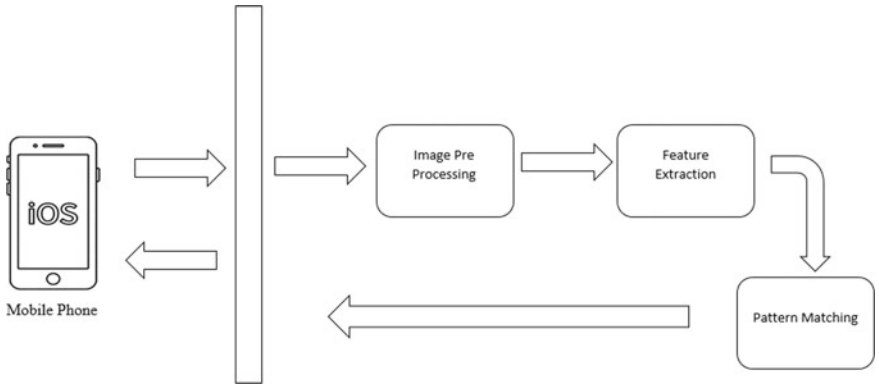


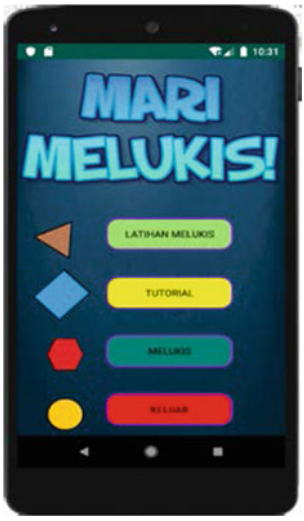

**Fig. 2** System architecture

## *3.1 User Interface*

There are three user interfaces designed for this system, which are the main page, selecting shapes, and drawing canvas. When users touch the application icon on the mobile phone, the system will go directly to this main page (Fig. 3). Two modules are provided in this application, namely tutorial and drawing. There are four option buttons provided to guide users to go through the applications, which are "Latihan Melukis" (Drawing Exercise), "Tutorial" (Tutorial), "Melukis" (Drawing), and "Keluar" (Exit). If users wish to test their drawing skills, they just need to touch the "Latihan Melukis" button.

Figure 4 shows the selecting shape interfaces. These interfaces will appear on the

**Fig. 3** Main page



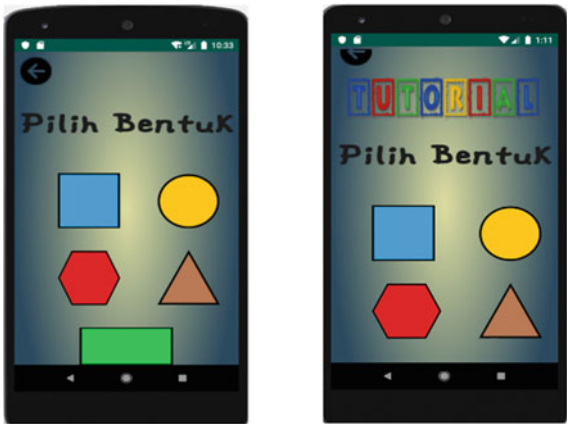



**Fig. 4** Selecting shape interfaces

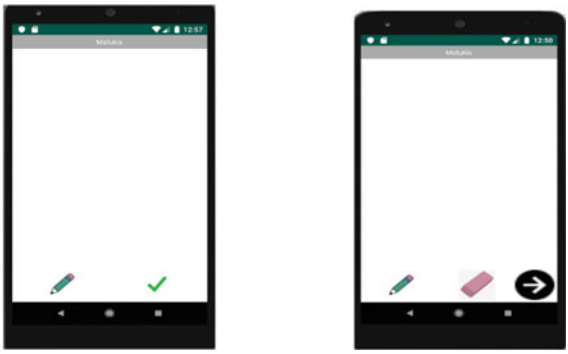


**Fig. 5** Drawing canvas interfaces

screen when users touch either "Latihan Melukis" or "Tutorial" button. There are five buttons provided representing five shapes. Users can choose any shape type they want to draw. Moreover, the interface offers a back button at the top left corner.

There are two different canvas types but with similar characteristics (Fig. 5). One canvas is to ensure the drawn shape is accurate by providing tools such as pencil and tick icons while the other canvas is to enable free drawing with the same tools as the previous canvas but with an eraser as an additional tool.

## *3.2 Real-Time Feedback Engine*

This system aims at identifying the best possibility to match the shape drawn by the children is achieved in this process by using pattern recognition, which will be

discussed in this section. Mainly, the purpose of the system purpose is to match the drawn shape pattern.

**Pattern Recognition**

Pattern Recognition is used in this system. The data will be collected among users. There are three pattern recognition phases, which are pre-processed analysis, feature extraction, and pattern matching.

1. Pre-processed Analysis
   In this phase, all shape type data that will be used in this system will be collected, such as every pixel line coordinate and the shape types. The coordinate is to calculate each pixel movement distance between x and y made by users when they draw each shape on the canvas provided. There will be many coordinates depending on how users draw. All coordinates are stored in an array.
2. Feature Extraction
   Feature extraction is a phase in which the shape features are extracted through every drawing movement performed by users. Besides, the collected coordinates can be used to identify the shape extraction by showing the calculation of x and y pixel distance. The extraction is divided into divisions, such as the top, middle, bottom, right, and left. This system uses an 8-connected neighbourhood Freeman chain code appending into a string and stored in an array.
3. Pattern Matching
   Every set of patterns, using the collected data, will be matched by the tools to deter-mine suitable matching patterns. Regular Expression tools are used to test whether the patterns are a match.

**Regular Expressions**

A regular expression or known as regex is a unique search pattern for the string text. A regular expression "engine" is a piece of software processing regular expression, trying to match the given string pattern. The regular expression has several syntaxes that can be used based on the logical creativity basically for the programmers. Figure 6 shows an example of the data taken.

Figure 6 shows an example of shape pattern data using regular expressions tools. Shape extraction features will be inserted in the test string, and the regular expression form will be used to make an expression matching the test string.

**Testing**

Testing was conducted to ensure the application functionality is satisfying and get feedback from users concerning the application usability for future use. The forms were distributed and answered by the respondents based on their experiences using the application. Table 1 shows the test case form, which is the test case for the project system evaluation.

The application results to match the shape patterns were validated whether they matched accurately with the users drawing. Table 2 shows the application validation results, while Fig. 7 shows the graph for each shape pattern average accuracy. Based

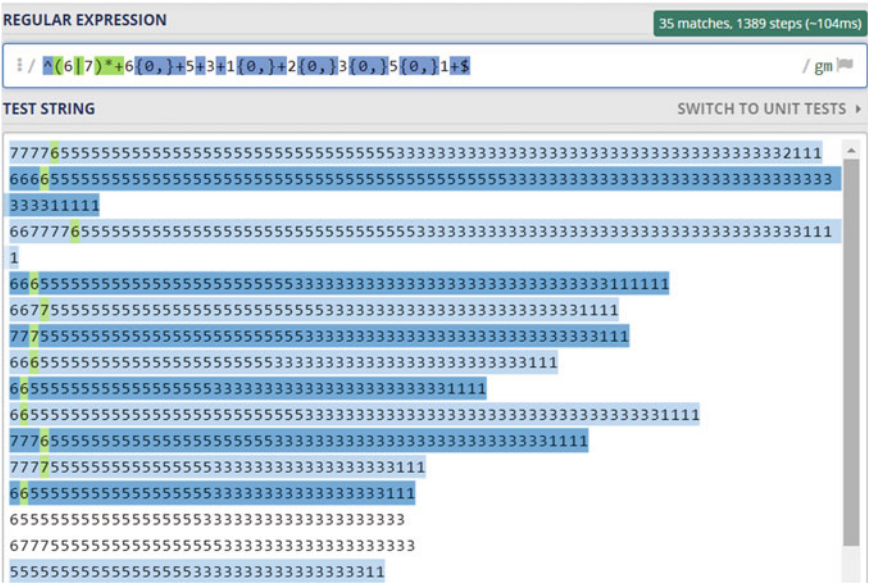



**Fig. 6** Regular expression tools

on Fig. 7, mostly, the average is between 80% and 100%. The average accuracy for drawing the circle and triangle is 100% while others are 80%. The circle and triangle achieved 100% accuracy due to the shape features that are easily matched by the regular expressions. Based on the users' feedback, the system will be further enhanced in the future.

Based on Fig. 7, mostly, the average is between 80% and 100%. The average accuracy for drawing the circle and triangle is 100% while others are 80%. The circle and triangle achieved 100% accuracies due to the shape features that are easily matched by the regular expressions. Based on the users' feedback, the system will be further enhanced in the future.

## 4 Conclusion

The regular expression in pattern matching was used in this project because the tool can match the patterns using expressions with the creativity of programmers. A set of data was collected, analysed, and tested using the regular expression to get suitable expressions to match the obtained data pattern. The application interface, which is the final project component, was designed and developed to complete the whole project. Functionality test was conducted to get the system performance accurate results. Based on the testing analysis, this project passed all the testing required. To conclude, the project can help children learning how to draw shapes accurately in a

**Table 1** Test case form

| Test case | Function | Description | Expected result | Actual result | Pass/Fail |
|---|---|---|---|---|---|
| 1 | "*Latihan Melukis*" button | Navigate users to the next page | Navigate users to selecting shapes | | Pass/Fail |
| 2 | "*Tutorial*" button | Navigate users to the next page | Navigate users to selecting shapes | | Pass/Fail |
| 3 | "*Melukis*" button | Navigate users to the next page | Navigate users to the drawing canvas | | Pass/Fail |
| 4 | "*Keluar*" button | Navigate users to exit the application | Users exit the application | | Pass/Fail |
| 5 | Shape button | Navigate users to the next page | Navigate users to the drawing canvas | | Pass/Fail |
| 6 | Left and Right Arrow buttons | Navigate users to the previous and next page | Navigate users to previous and next page | | Pass/Fail |
| 7 | Drawing Platform | The application enables users to draw | Users can draw on the canvas drawing platform | | Pass/Fail |
| 8 | Eraser and Pencil tools | The eraser can erase the drawing and pencil can draw an output on the drawing canvas | Users can draw using the pencil and can erase the drawing using the eraser | | Pass/Fail |

**Table 2** Application validation results

| Case validation | Scenario | Results (%) |
|---|---|---|
| 1 | Match pattern for drawing a Square | 80 |
| 2 | Match pattern for drawing a Circle | 100 |
| 3 | Match pattern for drawing a Hexagon | 80 |
| 4 | Match pattern for drawing a Triangle | 100 |
| 5 | Match pattern for drawing a Rectangle | 80 |

better way, especially for the children having difficulties knowing the shape types. For future work, this application should add sound effect to every of the shape that will make children to hear beside see (read) the images. Researchers said through Visual, Auditory, and Kinesthetic (VAK), children can have a better learning technique.

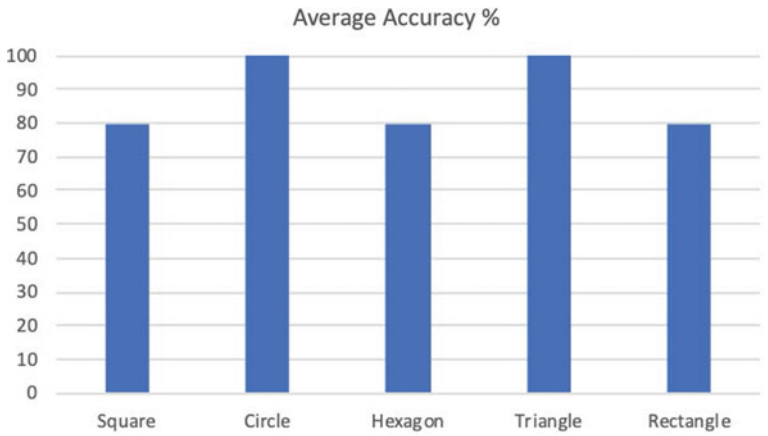



**Fig. 7** Average accuracy

**Acknowledgements** The authors would like to thank Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, for sponsoring this research.

## References

1. Romano E, Babchishin L, Marquis R, Fréchette S (2015) Childhood maltreatment and educational outcomes. Trauma Violence Abuse 16(4):418–437
2. Resnick I, Verdine BN, Golinkoff R, Hirsh-Pasek K (2016) Geometric toys in the attic? A corpus analysis of early exposure to geometric shapes. Early Childh Res Q 36:358–365
3. Verdine BN, Zimmermann L, Foster L, Marzouk MA, Roberta MG, Pasek KH, Newcombe N (2019) Effects of geometric toy design on parent–child interactions and spatial language. Early Child Res Q 46:126–141
4. Krig S (2014) Image pre-processing. In: Krig S (ed.) Computer vision metrics: survey, taxonomy, and analysis, pp. 39–83. Apress, Berkeley
5. Nixon M, Aguado A (2020) Feature extraction and image processing for computer vision, 4th edn. Academic Press of Elsevier, London
6. Ansari S (n.d.). https://www.geeksforgeeks.org/pattern-recognition-introduction/
7. Pistono F (2016). https://medium.com/@FedericoPistono/on-the-importance-of-pattern-recognition-6d7573d43595
8. Rouse M, Ansari S (n.d.). Retrieved from https://whatis.techtarget.com/definition/pattern-recognition
9. Brownlee J (2016). https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
10. Althobaiti H, Lu C (2017) A survey on Arabic optical character recognition and an isolated handwritten Arabic Character Recognition algorithm using encoded freeman chain code. In: Proceeding of 51st annual conference on information sciences and systems (CISS), Baltimore, MD, pp 1–6
11. Baji F, Mocanu M (2018) Chain code approach for shape based image retrieval. Ind J Sci Technol 11(3):1–17

12. Annapurna P, Kothuri S, Lukka S (2013) Digit recognition using Freeman chain code. Int J Appl Innov Eng Manag (IJAIEM) 2(8):2319–4847
13. Xia LY, Dai SG (2013) Circle and circular arc detection algorithm research based on Freeman chain code. In: IEEE 4th international conference on electronics information and emergency communication, Beijing, pp 230–233
14. Azmi AN, Nasien D (2014) Feature vector of binary image using Freeman chain code (FCC) representation based on structural classifier. Int J Adv Soft Comput Appl 6(2):30–35
15. Husain NH, Diah NM, Hanum HM (2017) The relevance of Freeman Chain Code for copying activities. In: 2017 IEEE conference on e-learning, e-management and e-services (IC3e), Miri, Serawak, pp 138–144
16. Azmi AN, Nasien D, Omar FS (2016) Biometric signature verification system based on Freeman chain code and k-nearest neighbor, Multimed Tools Appl 76:1–15
17. Liu P, Zhang J, Guo K (2015) A parking-lines recognition algorithm based on Freeman Chain Code. In: Proceeding of 7th international conference on intelligent human-machine systems and cybernetics, Hangzhou, pp 349–352

# Information Technology Students' Preferences on Blended Learning


**Choo-Kim Tan, Choo-Peng Tan, and Ng Shaun Wes**

**Abstract** The advancement of technology nowadays provides opportunities in education including the adoption of blended learning. The objective of this project was to examine Information Technology students' preferences on blended learning. Findings found that most students preferred to use it in Science subjects/courses, both inside the class and outside the classroom, and learn via online with a blended learning method as revision after the class. However, students of low and medium math anxious students had no significant difference in their preferences. It is hoped that the findings of the project will help researchers to customise the incorporation of blended learning tools in students learning as well help educators in planning and adopting blended learning, and benefit students in their learning. The system developers also should consider designing and develop more suitable systems/apps for students in learning Language subjects.

**Keywords** Information technology · Blended learning · Mathematics anxiety

## 1 Introduction

The rapid changes in this era which resulted from the advancement of sciences and technologies have provided many opportunities for development in most fields including education. It was found that many new educational tools, systems and applications such as Kahoot!, Prezi, Quizizz, Padlet, and many more have been developed and introduced in the educational settings. Their adoptions are widely found throughout the world. As such, there are many educational concepts being introduced

C.-K. Tan (✉) · C.-P. Tan · N. Shaun Wes
Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia
e-mail: cktan@mmu.edu.my

C.-P. Tan
e-mail: cptan@mmu.edu.my

N. Shaun Wes
e-mail: shaunngwes@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_9

in the educational word such as online learning, e-learning, flipped classroom, virtual classroom, blended learning, etc.

The advancement of technology nowadays provides opportunities in education including the adoption of blended learning. Blended learning is an approach that involves the combination of classroom and online education [1]. Oweis [2] stated that blended learning can be in the mixed forms of direct and indirect online learning. It is conducted using the internet and intranet. Furthermore, the indirect learning takes place concurrently in the traditional classes. Ceylan [3] defined blended learning as a computer-mediated instructional strategy that involves and uses technology and emphases student-teacher relationship. Students' achievement, engagement and independence in learning improved with its adoption too.

Twelve most common types of blended learning were identified by Teachthought [4]. These include Station Rotation Blended Learning, Lab Rotation Blended Learning, Remote Blended Learning (also referred as Enriched Virtual), Flex Blended Learning, the 'Flipped Classroom' Blended Learning, Individual Rotation Blended Learning, Project-Based Blended Learning, Self-Directed Blended Learning, Inside-Out Blended Learning, Outside-In Blended Learning, Supplemental Blended Learning, and Mastery-Based Blended Learning. Which type of blended learnings to adopt will depend on many factors such as students' learning styles, nature of the subjects, the preferences of the teachers and students, etc.

Every individual has his/her own preferences of learning styles that may affect his/her learning outcomes and influence his/her ability to attain information, to interact with peers and teachers, and also the participation in learning experiences [5]. A study found that blended learning differs according to students' learning styles [5]. The differences can be either their preferences to have blended learning for different subjects/courses, learning settings, or on different devices. Blended learning can be held in the classroom during class hours or outside the class [6]. In addition, blended learning can also be adopted as preparation before a class and review after a class [6]. It was found that there was no significant difference between outside-the-classroom blended class with the inside-the-classroom blended class [6]. Shantakumari [7] found that students' perceptions of blended learning differed with courses.

Furthermore, studies also found that the important factors for the effectiveness of blended learning include students' characteristics and the blended learning design features [8]. Thus, educators should be aware of the relationship between the students' characteristics, design features and learning outcomes.

There are benefits of implementing blended learning in educational institutions. Kintu, Zhu and Kagambe [8] stated that students' levels of knowledge construction increase in blended learning as the analytical skills are created in them. They found that students established assessing ability and the ability in evaluating knowledge sources analytically. In addition, blended learning helped students to develop leadership skills and confidence, gained new theoretical knowledge and needs analysis as well as the application of theoretical knowledge to practical educational projects [9]. Oweis [2] found that students who were taught with a blended learning method for English showed high motivation and performed significantly better than students who were taught using the traditional teaching method.

This project aims to examine students' preferences on blended learning, particularly Information Technology students. It is hoped that the findings of the project will help educators to implement blended learning effectively.

## 2 Methodology

### *2.1 Sample*

The project was carried out on a convenient sample on a small scale of 70 undergraduate students at a Malaysia private university in a period of one semester. These students enrolled for the Information Technology Program. Students' ages ranged from 18 to 27 years old. The sample consisted of multiracial (Malay, Chinese, Indian, etc.).

### *2.2 Instrument*

The project involved a set of questionnaires which consisted of 2 sections (Sections A and B).

1. Section A: This section is to collect students' general information such as gender, age, nationality, race, course and etc.
2. Section B: This section consists of questions on students' preferences on blended learning such as the preferred fields/subjects to implement blended learning and preferred time to conduct blended learning.

The questionnaires were validated by a panel of experts in education. The reliability test was run and generated a Cronbach Alpha Coefficient of .729, which indicated that the instruments generated good reliability coefficient and implied that it has internal consistency.

The questionnaire was given to the respondents after they experienced the blended learning approach. The data collected was analysed using SPSS for descriptive statistics, t-test, and ANOVA. 5% significance level was used to run the tests.

Furthermore, qualitative data was collected via interview with students in order to gather information to compliment the quantitative data. Questions such as "Do you like to learn with blended learning?", "Why do you prefer to learn with blended learning in this Mathematics/Information Technology/Languages?", "Why do you like to learn online inside/outside the classroom?", etc.

### *2.3 Procedure*

The blended learning approach adopted was online learning mixed with face-to-face learning in the classroom. In this project, the blended learning tools used were Multimedia Learning System (MMLs), Quizzes, Kahoot!, Prezi, Youtube and G-Suite Education (such as Google Classroom).

The MMLs is a system developed by the university and used in the university. It allows the uploading of lecture notes, tutorial questions, exercises, assignments/projects, quiz questions and making announcements. In addition, MMLs permits the conduct of quizzes and tests online which students attempt the questions within a stipulated date and time duration. The reports of the quizzes and tests will be generated for lecturers. The Discussion Board is another function integrated in MMLs.

For this project and with the blended learning approach, all the lecture notes and tutorial questions were uploaded into the MMLs for students. Also, the lecture materials were delivered via Prezi in which the lecturers designed and created their own notes presentation using Prezi. Thus, besides the face-to-face delivery of knowledge and concepts, students learnt through MMLs and Prezi. The assignments and projects were also uploaded in MMLs. Furthermore, online quizzes and tests were given to students via MMLs, Quizizz and Kahoot!. However, students were playing online games with Quizizz and Kahoot!. Videos were watched from Youtube. Google Classroom was used for the assignments/project's submission. Discussion among students and with lecturers were conducted either face-to-face or via MMLs.

The online learning was used to support students' activities inside the classroom, outside the classroom, before the class as class preparation and after the class as revision.

## 3 Findings and Discussions

This section discusses the findings generated from the analysis. The findings on students' preferred field or subject to learn with blended learning will be presented and discussed first, then followed by students' preferences of having blended learning in classroom or outside the classroom, and finally students' preferences of having blended learning before the class as class preparation or after the class as revision will be discussed.

Students in the sample was asked to give their preferences of implementing blended learning on the 3 fields which they enrolled in the semester, i.e. Mathematics, Information Technology subject and Languages. It was found that 30 (42.86%) students preferred to have blended learning in Mathematics, another 30 (42.86%) students preferred that blended learning be adopted for the Information Technology field, while 10 (14.28%) students have the preference on Languages subjects (Table 1).

**Table 1** Frequency and percentage of the field or subject students' would prefer to learn with blended learning based on students' mathematics anxiety levels

| | Anxiety level | Mathematics | ICT | Languages | Total |
|---|---|---|---|---|---|
| In what field/subject you would prefer to have blended learning? | Low | 5 (41.67%) | 4 (33.33%) | 3 (25.00%) | 12 (100.00%) |
| | Medium | 25 (43.10%) | 26 (44.83%) | 7 (12.07%) | 58 (100.00%) |
| Total | | 30 (42.86%) | 30 (42.86%) | 10 (14.28%) | 70 (100.00%) |

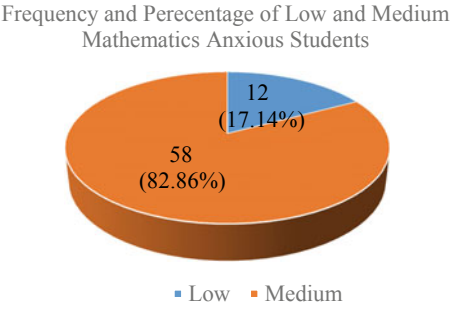



**Fig. 1** Frequency and percentage of students' based on mathematics anxiety levels

Further analysis was done on categorising students into three different groups based on their mathematics anxiety levels. Students were grouped in such a way because this paper was a part of the whole project on reducing students' mathematics anxiety levels using blended learning. Students' mathematics anxiety levels were obtained from the questionnaire. These data enabled the categorisation into high, medium and low anxious groups. Findings found that there were only medium and low maths anxious students but no high anxious students.

Among 70 students, the majority of them are medium math anxious students. Figure 1 shows that only 12 (17.14%) low math anxious students while 58 (82.86%) are medium math anxious students.

When students were asked about their preferred subject to learn with blended learning, among the 12 low math anxious students, 5 preferred Mathematics, 4 preferred Information Technology and 3 preferred Languages. From 58 medium math anxious students, almost the same number of students preferred to have blended learning in learning Mathematics (25 students/43.1%) and Information Technology (26 students/44,83%). While there are only 7 medium math anxious students preferred to learn Languages with blended learning.

Table 2 reports that there is no significant difference between low- and medium-math anxious students on their preferences of field or subject to learn with blended learning, $t(68) = .640$, $p > .05$.

The results showed that there are different preferences on the subjects by students which aligned with Shantakumari [7] that students' perceptions of blended learning differed with courses. Most students prefer blended learning in Sciences than

**Table 2** Compare of means on the field or subject students' would prefer to learn with blended learning based on students' mathematics anxiety levels

| | Anxiety Level | N | M | SD | MD (SD) | t-value | df | p-value |
|---|---|---|---|---|---|---|---|---|
| In what field/subject you would prefer to have blended learning? | Low | 12 | 1.83 | 0.835 | 0.144 | 0.640 | 68 | 0.524 |
| | Medium | 58 | 1.69 | 0.681 | (0.225) | | | |

Languages subjects. This finding contradicts with Oweis's [2] finding that students who learnt English with blended learning were highly motivated. Students of this project feedback that they prefer blended learning in Science subjects than Languages subjects because they found that most abstract concepts could be easily understood through the systems/apps. In addition, the sample comprised Information Technology students, thus they might have the preference in Sciences subjects. However, further findings found that students of different anxiety levels showed no significant difference in their preferences with subjects. This showed that blended learning method benefits students regardless of adopting it in any subjects and anxiety levels, which can be proven from Oweis's study [2] that students who were taught English with a blended learning method showed high motivation and performed significantly better than their counterparts who learned with the traditional teaching method and from this project that students who learned Sciences subjects and Languages subjects have no significant difference in their preferences when the blended learning method was used.

Furthermore, students in the sample were asked to give their preferences of having blended learning in the classroom or outside the classroom or both. Most of the students (27 students) preferred to have blended learning which online learning is used in both inside and outside the classroom. The least number of students, 20 students, liked to have blended learning which online learning is used outside the classroom, whereas 23 students preferred to have blended learning which online learning is used inside the classroom (Table 3).

Findings in Table 4 also show that among the 12 low math anxious students, there are an equal number of students preferred to have blended learning which online learning is used either inside the class or outside the classroom, which is 5 students respectively. Only 2 low math anxious students liked to learn via online inside the

**Table 3** Frequency and percentage of students' preferences of having blended learning in the classroom or outside the classroom

| | In the classroom | Outside the classroom | Both | Total |
|---|---|---|---|---|
| What if your preference of having blended learning, i.e. in the classroom or outside the classroom? | 23 (32.86%) | 20 (28.57%) | 27 (38.57%) | 70 (100.00%) |

**Table 4** Frequency and percentage of students' preferences of having blended learning in the classroom or outside the classroom based on students' mathematics anxiety levels

| | Anxiety level | In the classroom | Outside the classroom | Both | Total |
|---|---|---|---|---|---|
| What if your preference of having blended learning, i.e. in classroom or outside the classroom? | Low | 5 (41.67%) | 5 (41.67%) | 2 (16.67%) | 12 (100.00%) |
| | Medium | 18 (31.03%) | 15 (25.86%) | 25 (43.10%) | 58 (100.00%) |
| Total | | 23 (32.86%) | 20 (28.57%) | 27 (38.57%) | 70 (100.00%) |

**Table 5** Compare of means on students' preferences of having blended learning in the classroom or outside the classroom based on students' mathematics anxiety levels

| | Anxiety Level | N | M | SD | MD (SD) | t-value | df | p-value |
|---|---|---|---|---|---|---|---|---|
| What if your preference of having blended learning, i.e. in the classroom or outside the classroom? | Low | 12 | 1.75 | .754 | −0.371 | −01.385 | 68 | 0.170 |
| | Medium | 58 | 2.12 | 0.860 | (0.268) | | | |

classroom and also outside the classroom. In contrast, the majority of the medium math anxious students (25 students or 43.10%) would prefer to have blended learning where they learn through online both inside and outside the classroom. It is reported that 18 and 15 students preferred to learn via online with blended learning method either in the class or outside the classroom respectively.

As shown in Table 5, there is no significant difference between low- and medium-math anxious students on their preferences to have blended learning which online learning is implemented inside the classroom or outside the classroom, $t(68) = -1.385$, $p > .05$.

Even though there were students who liked to use online learning either inside the classroom or outside the classroom, most students liked to learn via online in both inside and outside the classroom with a blended learning method. Students suggested that online learning is very convenient for them to learn at any time they like, thus they prefer to learn online both inside and also outside the classroom. Even though, Miyaji's [6] study found that there was no significant difference between outside-the-classroom blended class with the inside-the-classroom blended class, the findings of this project found further findings which indicating no significant difference among students of different anxiety levels on their preferences to have online learning with blended learning method inside the classroom or outside the classroom.

**Table 6** Frequency and percentage of students' preferences of having blended learning before class as class preparation or after class as revision based on students' mathematics anxiety levels

| | Anxiety level | Before class as class preparation | After class as revision | Both | Total |
|---|---|---|---|---|---|
| Would you like to prepare yourself with blended learning before or after the class? | Low | 1 (8.33%) | 9 (75.00%) | 2 (16.67%) | 12 (100.00%) |
| | Medium | 13 (22.41%) | 31 (53.45%) | 14 (24.14%) | 58 (100.00%) |
| | Total | 14 (20.00%) | 40 (57.14%) | 16 (22.86%) | 70 (100.00%) |

Students were also asked to indicate their preference either to have online learning with blended learning method before the class, i.e. students learn the subject materials online before they attend the class (face-to-face) as the preparation before the class, or to use online learning for revision/enhancement after they attended the class (face-to-face), or both.

Table 6 shows the findings of students' preferences of having online learning before the class as the class preparation or after the class as the revision of teaching materials/enhancement of knowledge, or both.

Majority of the students, 40 students (57.14%), preferred blended learning by having online learning for their revision after the face-to-face learning in the classroom. Preparation for class via online learning before face-to-face learning recorded the lowest number of students, which is only 14 students (20.00%). There are 16 students who like to have blended learning when the online learning is used as class preparation before the face-to-face class and as revision after the face-to-face learning.

Among the 40 students who preferred blended learning by having online learning for their revision after the face-to-face learning in the classroom, there are 9 (75.00%) of the low math anxious students and 31 (53.45%) of the medium math anxious students. This is followed by 2 (16.67%) of the low math anxious students and 14 (24.14%) of the medium math anxious students who liked to have blended learning by using online learning as both class preparation before the face-to-face learning in the class and revision after the face-to-face learning. Only 1 (8.33%) low math anxious student and 13 (22.41%) medium math anxious students who liked the blended learning when they learn the materials via online before the class as class preparation.

Table 7 reported that there is no significant difference between low- and medium-math anxious students on their preferences of having blended learning when online learning is used as class preparation before the class or as revision after the class, $t(68) = .314$, $p > .05$.

As stated by Miyaji [6] that blended learning can be adopted as preparation before a class and review after a class. This project found that most of the students preferred

**Table 7** Compare of means on students' preferences of having blended learning before class as class preparation or after class as revision based on students' mathematics anxiety levels

| | Anxiety level | N | M | SD | MD (SD) | t-value | df | p-value |
|---|---|---|---|---|---|---|---|---|
| Would you like to prepare yourself with blended learning before or after the class? | Low | 12 | 2.08 | 0.515 | 0.066 | 0.314 | 68 | 0.754 |
| | Medium | 58 | 2.02 | 0.688 | (0.210) | | | |

to learn via online after the face-to-face class. Students liked to do revision using online materials after their lecturers have delivered knowledge in the classroom. Their knowledge is enhanced by reading and doing more online activities when they have understood the concept/theories delivered by the lecturers in the classroom. Students of different anxiety levels have no significant preferences of having online learning as class preparation before the class and as revision after the class. They enjoyed online learning either as preparation or revision.

## 4 Conclusion

Technology helps in students' learning. This project discusses the students' preferences associated with blended learning, an approach which involves technology in improving teaching and learning. Basically, most research found that blended learning is an effective teaching and learning approach. It is a mixture of online learning and face-to-face learning, which is not 100% online learning or 100% face-to-face learning in the classroom, in which either approach may make the learning uninteresting. Blended learning allows a variety of online and face-to-face activities for students. It diversifies the teaching and learning.

It can be concluded from this project that when blended learning is adopted, most students prefer to use it in Science subjects/courses, both inside the class and outside the classroom, and learn via online with a blended learning method as revision after the class. Therefore, the findings of this project recommend that to teach Science students, it is preferable to adopt blended learning in Sciences subjects than Language subjects. Also, it is suggested that when planning lessons, lecturers should consider more online activities and appropriate online activities to be used inside and outside the classroom. The online activities for after the classroom should be more to the revision activities and materials. However, students of low and medium math anxious students had no significant difference in their preferences. Thus, it is suggested that blended learning could be a suitable learning approach for all students regardless of their math anxiety levels.

Finally, it is hoped that the findings of this project could help researchers to customise the incorporation of blended learning tools in students' learning as well help educators in planning and adopting blended learning, and benefit students in

their learning. Furthermore, the system developers should consider designing and developing more suitable systems/apps for students in learning Language subjects.

Future research will be done on examining the effects of using blended learning to reduce students' mathematics anxiety levels. A system will be developed and to be used as a technology tool in this future research.

**Acknowledgements** We would like to thank Multimedia University for the mini fund to support the conduct of this project.

## References

1. TeachThought Staff (2020) The Definition of Blended Learning. Teachthought. https://www.teachthought.com/learning/the-definition-of-blended-learning/. Accessed May 2020
2. Oweis TI (2018) Effects of using a blended learning method on students' achievement and motivation to learn English in Jordan: a pilot case study. Educ Res Int 2018:1–7. https://doi.org/10.1155/2018/7425924
3. Ceylan VK (2017) Effect of blended learning to academic achievement. https://doi.org/10.14687/jhs.v14i1.4141
4. TeachThought Staff (2019) 12 Of The Most Common Types of Blended Learning. Teachthought. https://www.teachthought.com/learning/12-types-of-blended-learning/. Accessed 23 May 2020
5. Akkoyunlu B, Soylu MY (2008) A study of student's perceptions in a blended learning environment based on different learning styles. Educ Technol Soc 11(1):183–193
6. Miyaji I (2011) Comparison between effects in two blended classes Which E-learning is used inside and outside classroom. US-China Educ Rev 8(4):468–481
7. Shantakumari NS (2015) Blended learning: the student viewpoint. Ann Med Health Sci Res 5(5):323–328
8. Kintu MJ, Zhu C, Kagambe E (2017) Blended learning effectiveness: the relationship between student characteristics, design features and outcomes. Int J Educ Technol High Educ 14(7) (2017). https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-017-0043-4
9. Namyssova G, Tussupbekova G, Helmer J, Malone K, Afzal M, Jonbekova D (2019) Challenges and benefits of blended learning in higher education. Int J Technol Educ (IJTE) 2(1):22–31

# Improved Facial Recognition Algorithms Based on Dragonfly and Grasshopper Optimization


**Dyala Rasheed Ibrahim, Je Sen Teh, and Rosni Abdullah**

**Abstract** In this paper, we investigate two relatively new optimization algorithms in facial recognition, the grasshopper optimization algorithm (GOA) and binary dragonfly algorithm (BDA) which had the best performance out of 13 optimization algorithms that were compared. We investigate the effectiveness of both optimization algorithms alongside two classifiers, k-nearest neighbor (KNN) and support vector machine (SVM). Performance evaluation of the four combinations, BDA-KNN, BDA-SVM, GOA-KNN and GOA-SVM, indicate near-ideal recognition rates, with the GOA variants slightly outperforming their BDA counterparts. When compared to other recently proposed facial recognition approaches, the proposed algorithms depict improved accuracy.

**Keywords** Biometrics · Binary dragonfly algorithm · Classification · Facial recognition · Grasshopper algorithm · Optimization algorithm

## 1 Introduction

Facial recognition (FR) has many practical applications due to its advantages such as uniqueness, immutability, social acceptance, ease of use and low cost [1]. It is a nonintrusive method for identifying or verifying individuals. FR algorithms involve training classifiers using facial features. Unfortunately, there are many redundant, irrelevant features negatively affect the performance of FR approaches. Approaches such as binary pattern (LBP) can be used to extract local spatial patterns as opposed

D. R. Ibrahim · R. Abdullah
National Advanced Ipv6 Centre, Universiti Sains Malaysia, Penang, Malaysia
e-mail: ahmaddyalahdeeb@gmail.com

R. Abdullah
e-mail: rosni@usm.my

J. S. Teh (✉) · R. Abdullah
School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
e-mail: jesen_teh@usm.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_10

to global features [2, 3]. It is a feature descriptor for facial expression representation. The main advantages of LBP are tolerance against illumination changes and its computational simplicity [4]. For further performance improvements, feature selection methods can be employed to reduce feature space dimensionality. Feature selection attempts to solve the problem of redundant, irrelevant, and inaccurate features, and can be performed with the aid of various optimization algorithms. These algorithms can include ant colony optimization [5], particle swarm optimization [6], bacteria foraging optimization [7] and firefly algorithm [8].

FR can be implemented using classifiers which include artificial neural networks (ANNs), support vector machine (SVM) and the k-nearest neighbor (KNN) algorithm. These are machine learning approaches that are commonly used for pattern recognition. Many researchers have shown that KNN and SVM outperform other classifiers for FR purposes [9–11]. KNN is a simple, efficient, reliable, and computationally efficient algorithm for FR [12] whereas SVM is a machine learning approach widely used in image processing applications. KNN has high recognition rate and can quickly identify items from a large dataset [10]. In terms of facial recognition, KNN leverages upon distance metrics to identify the closest person from the dataset [11, 13]. SVM is an effective discriminative classifier that developed by Philips [14]. The input to SVM is a set $X, Y$ of labeled training data, where $X$ is the data and $Y = [-1, 1]$ is the label. The output of an SVM algorithm is a set of $N$ support vectors. The main advantage for SVM is stability, whereby bit changes in the data do not greatly affect the hyperplane, leading to a stable model [15, 16]. SVM can be used to develop classifier or regression models. For FR, SVM attempts to generate a decision surface that separates dissimilarities between images of the same and the images of different individuals [11].

Various FR algorithms have been proposed in recent literature, all with the goal of maximizing recognition rate by adopting a variety of techniques. Gao and Lee's approach is based on scale-invariant feature transform (SIFT), which is a method to extract local features [17]. The experimental results shown that average performance of 95% when tested on the FERET dataset. Agarwal and Bhanot's approach involved identifying the center hidden neuron layers of a radial basis function neural network for the purpose of facial recognition [8]. They used the firefly optimization algorithm as feature selection method. Experimental results showed decent recognition rates for various databases as ORL (97.75%), Yale (99.83%), AR (93.15%) and LFW (60.50%). Zhu and Xue presented a novel approach called random subspace method for FR [18]. The tensor subspace approach was used for feature selection to achieve a recognition rate of 98.32%. Lu et al. used a sparse representation method using rank decomposition to get a robust recognition rate of 96% [19].

Other researchers developed FR methods to address issues of high-dimensional features and the multitude of variations available in face images. One method uses GOA to extract relevant features from the high dimensional feature vectors [20]. Their experiments on the ORL dataset led to an accuracy of 91.5%. Sasirekha and Thangavel proposed novel FR algorithm based on KNN with particle swarm optimization (PSO) [21]. LBP and PSO were used to extract and select features respectively, leading to a best-case accuracy of 97.41%. Maheshwari et al. developed an

FR approach based on local directional pattern, a feature extraction method [14]. Then, genetic, and differential optimization algorithms were used as a feature selection method to eliminate the irrelevant features. Finally, SVM used to classify the identity of facial images. Their experimental analysis showed that differential evolution outperforms genetic algorithm. Gupta and Goel developed a FR approach that extracts features using a Gabor filter [22]. Principal component analysis (PCA) was then used for feature selection for dimension reduction. A modified version of the artificial bee colony (ABC) is then used on the feature vectors to search for the best match for a test image in a given database, achieving an accuracy of 97%. Abd et al. proposed an FR approach also based on the Gabor filter for feature extraction followed by feature selection by grey wolf optimization (GWO) algorithm. By training a KNN classifier, a recognition rate 97% was achieved on the Yale dataset [23]. The FR approach by Kumar, based on PCA and bat optimization algorithm depicted a recognition accuracy of 96% when tested on the Yale database [12].

More recently, Aro et al. proposed an FR algorithm based on enhanced gabor filters and the ant colony optimization algorithm [24]. The proposed method aimed to solve the high dimensionality problem of gabor filters that lead to low performance and high time complexity. The ant colony optimization algorithm was used to remove noisy, redundant and irrelevant gabor features. They achieved an accuracy of 97.14% and 95.71% using the Malahanobis and Chebyshev classifiers, respectively. Benamara et al. proposed a multispectral face recognition method using random feature selection and PSO-SVM [25]. The proposed method solved the problem of intra-variation conditions which negatively affects the performance of FR systems by using both infrared and visible spectra. A new feature selection algorithm was introduced that reduces the feature space dimensionality to be suitable for real time applications.

Eleyan proposed a PSO metaheuristic algorithm as a feature selection method for face recognition systems that reduces the dimensionality of extracted feature vectors [26]. Experimental analysis was executed by using two well-known face databases. Performance of the PSO approach in terms of accuracy, specificity and sensitivity depicted high performance as compared to other algorithms such as principal component analysis (PCA). Malhotra and Kumar proposed an optimized facial recognition approach that combines DCT and PCA to extract the features that led to a high recognition accuracy of 96.5% [27]. Cuckoo search was used in the feature selection stage to remove irrelevant features. Král et al. proposed another face recognition system based on an improved local binary patterns (LBP) approach [28]. In the proposed approach, the enhanced LBP considers more pixels and different neighborhoods while computing the features. The proposed approach was evaluated using UFI and FERET face datasets, depicting improved performance as compared to other state-of-the art approaches. Table 1 shows the summary of the related work.

In this paper, we investigate the use of two relatively new optimization algorithms in facial recognition. We select these algorithms after studying different 13 optimization algorithms from the perspectives of accuracy and time complexity when used for feature selection. Based on our experiments, the binary dragonfly algorithm (BDA) and grasshopper optimization algorithm (GOA) outperformed their peers in both

**Table 1** Summary of related work

| Reference | Dataset | Method | Best results (%) |
|---|---|---|---|
| [8] | AR, LFW, ORL, Yale | Firefly, neural network | 99.83 |
| [37] | FERET, ORL, Yale | Gabor filter, genetic, SVM | 99.30 |
| [28] | UFI, FERET | – | 98.5 |
| [25] | Visible and CSIST | PSO | 98 |
| [26] | ORL, PUT | PSO | 98 |
| [21] | ORL | PSO, KNN | 97.41 |
| [24] | ORL, AFI | Ant colony optimization | 97.14 |
| [22] | ORL | ABC, PCA | 97.00 |
| [23] | Yale | Gabor filter, GWO, KNN | 97.00 |
| [27] | ORL | Cuckoo search | 96.5 |
| [12] | Yale | Bat, PCA | 96.00 |
| [19] | AR, FERET, ORL | Rank decomposition, least squares | 96.00 |
| [17] | FERET | SIFT, correspondence learning | 95.00 |
| [38] | ORL | Genetic, differential, SVM | 95.00 |
| [18] | AR, CMU PIE, Yale | Random subspace | 93.14 |
| [20] | ORL | GOA | 91.50 |

aspects. Both optimization algorithms are used for feature selection prior to training KNN and SVM classifiers. We denote the four FR approaches as BDA-KNN, BDA-SVM, GOA-KNN and GOA-SVM. The proposed FR algorithms depict a desirable performance in terms of both recognition rate and time complexity, outperforming other recently proposed FR algorithms.

The remainder of this paper is organized as follows: Sect. 1 discusses related work in FR, followed by Sect. 2 which investigates 13 optimization algorithms for feature selection. Section 3 then describes four of the proposed FR approaches whereas Sect. 4 provides experimental analysis of those methods. Finally, the paper concludes with some final remarks in Sect. 5.

## 2 Optimization Algorithms

### *2.1 Binary Dragonfly Algorithm*

The dragonfly algorithm (DA) is a relatively new optimization algorithm based on swarm intelligence proposed in 2016 [29]. There are many versions of DA such as BDA, multi-objective dragonfly algorithm and single-dragonfly algorithm. The relevant parameters for BDA are listed below, where N is the number of neighboring individuals, $X_i$, $X_j$, $X^+$, $X^-$ denote the positions of the current individual,

$j$ th individual, food source and enemy respectively, and $t$ denotes the number of iterations,

$$Separation{:}S_i = -\sum_{j=1}^{N}(X_i - X_j), \tag{1}$$

$$Alignment{:}A_i = \frac{\sum_{j=1}^{N} Vj}{N}, \tag{2}$$

$$Cohesion{:}C_i = \frac{\sum_{j=1}^{N} Xj}{N} - X_i, \tag{3}$$

$$Attraction{:}F_i = X^+ - X_i \tag{4}$$

$$Distraction{:}E_i = X^- + X_i, \tag{5}$$

To update the position of dragonflies in a search space and formulate their movements, two vectors are considered, the step vector $\Delta X$ and position, $X$. The step vector denotes the direction of dragonfly movement which can be calculated as

$$\Delta X_{t+1} = (sS_i + aA_i + cC_i + fF_i + eE_i) + w\Delta X_t. \tag{6}$$

After calculating the step vector, the position vectors are calculated as

$$X_{t+1} = X_t + \Delta X_{t-1} \tag{7}$$

Then, to enhance the randomness of the dragonflies,

$$X_{t+1} = X_t + \left(0.01X_t \times \frac{r_1 \times \alpha}{|r_2|^{\frac{1}{\beta}}}\right), \tag{8}$$

where $r_1$, $r_2$ denote two random numbers in [0,1], $\beta = 1.5$ and α is calculated as

$$\alpha = \left(\frac{\Phi(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Phi\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}}\right)^{\frac{1}{\beta}}, \tag{9}$$

where $\Phi(x) = (x-1)!$. Finally, the transfer function is used to calculate the probability of the dragonflies changing positions,

$$T(\Delta x) = \left| \frac{\Delta x}{\sqrt{\Delta x^2 + 1}} \right|. \tag{10}$$

To update the position of search agents in binary search spaces,

$$X_{t+1} = \begin{cases} \neg X_t, r < T(\Delta X_t + 1) \\ X_t, r \geq T(\Delta X_t + 1) \end{cases}, \tag{11}$$

where $r$ denotes to a number in the interval of [0,1]. The BDA algorithm considers all the dragonflies as one swarm and simulate exploration/exploitation by adaptively tuning the swarming factors ($s$, $a$, $c$, $f$, and $e$) as well as the inertia weight ($w$). The pseudocode of BDA is shown in Algorithm 1.

**Algorithm 1** Dragonfly Algorithm

```
Initialize the population of dragonflies X_i where i = {1,2,...,n}
Initialize step vectors ΔX_i where i = {1,2,...,n}
while the end condition is not satisfied
  Calculate the objective values of all dragonflies
  Update the food source and enemy
  Update w, s, a, c, f, and e values
  Calculate S, A, C, F, and E values (Eq. 1-5)
  Update step vectors (Eq. 6-9)
  Calculate probabilities of changing position for all dragonflies (Eq. 10)
  Update position vectors (Eq. 11)
end while
```

## 2.2 *Grasshopper Optimization Algorithm*

Grasshopper optimization algorithm (GOA) is a new optimization algorithm proposed by Saremi et al. in 2017 [30]. The multi-objective version of the grasshopper algorithm was later proposed in 2018 proposed by Mirjalini [31]. As its name implies, GOA is inspired from the behavior of grasshoppers. It is generally used to search for optimal solutions to constrained and unconstrained problems [32]. The pseudocode of GOA is shown in Algorithm 2. The position of the $i$th grasshopper, $X_i$ is calculated as

$$X_i = S_i + G_i + A_i, \tag{12}$$

where $S_i$ is the social interaction, $G_i$ is the gravitational force on the $i$th grasshopper and $A_i$ is the wind advection. Social interaction is the main parameter that dictates the grasshoppers' movement which can be calculated as

$$S_i = \sum_{j=1, j\neq i}^{N} s(dij)\widehat{d_{ij}}, \tag{13}$$

where, $N$ is the number of grasshoppers, $d_{ij}$ is the distance between the *ith* and *jth* grasshoppers, $\widehat{d_{ij}}$ is a unit vector from the *ith* to the *jth* grasshopper, and $s$ is a where, $N$ is the number of grasshoppers, $d_{ij}$ is the distance between the *ith* and *jth* grasshoppers, $\widehat{d_{ij}}$ is a unit vector from the *ith* to the *jth* grasshopper, and $s$ is a function that represents social attraction. These parameters are defined as

$$d_{ij} = \left|x_j - x_i\right|, \tag{14}$$

$$\widehat{d_{ij}} = \frac{x_j - x_i}{d_{ij}}, \tag{15}$$

$$s(r) = fe^{-\frac{r}{l}} - e^{-r}, \tag{16}$$

respectively, where $f$ and $l$ are the attraction intensity and the attractive length scale respectively, and $x_i$ represents the *ith* grasshopper within the entire population. The final mathematical model of the grasshopper position in the $d$th dimension is described as

$$sX_i^d = c\left(\sum_{j=1, j\neq i}^{N} c\frac{ub_d - lb_d}{2} s\left(d_{ij}\right)\widehat{d_{ij}}\right) + \widehat{T_d}, \tag{17}$$

where $ub_d$, $lb_d$ and $\widehat{T_d}$ are the upper bound, lower bound and best solution found so far, respectively. $c$ is a control parameter to modify the behavior of exploitation and exploration and can be calculated as

$$c = c_{max} - l\frac{(c_{max} - c_{min})}{L}, \tag{18}$$

where $c_{max} = 1$, $c_{min} = 0.00001$, $l$ and $L$ are the maximum value, minimum value, current iteration and maximum number of iterations, respectively.

**Algorithm 2** Grasshopper Algorithm

```
Initialize the population of grasshoppers X_i where i = {1,2,...,N}, each with d dimensions
Initialize c_max, c_min and L
Set the best current solution as the target vector, T̂_d
while l < L
    Calculate c (Eq. 12)
    For each grasshopper, I do
          Calculate d_ij (Eq. 14) and update X_i^d (Eq. 17)
          Apply boundary checks on each solution
      End for
end while
```

### *2.3 Comparison of Optimization Algorithms*

Prior to selecting BDA and GOA to be used in our work, we performed a comparison of 13 optimization algorithms according to 12 test functions to determine their accuracy and efficiency for feature selection purposes. The 12 test functions used for comparison include Raster, Ackley, Camel3, Dejong5, Levy, Sphere, Rosen, Griewank, Zakharov, Schaffer2, Rothyp and Shubert [33]. Experiments were performed using MATLAB 2018 on an Intel Core-i5 CPU with 2 GB RAM. The experiments were executed 1000 times before the accuracy results (cost function) and time taken (in seconds) for each execution are noted, where for both measures, a lower value is desired. Search area dimensions between 10, 20 and 30 were used, with the lower and upper bounds of $10 \in [-5, 5]$, $20 \in [-10, 10]$ and $30 \in [-15, 15]$. Only the unimodal category (single solution problems) is used to determine the best algorithm for feature selection. The results are tabulated in Table 2, where the dragonfly and grasshopper optimization algorithms outperformed their peers in both metrics.

## 3 Proposed Method

In the proposed work, features of the human face are first extracted using uniform LBP (ULBP). Features are the significant characteristics from a face image which may be its shape, texture, or context. Relevant features are then selected by using BDA and GOA to train two classifiers, KNN and SVM. Classifiers trained using features selected by BDA are denoted as BDA-KNN and BDA-SVM whereas the

**Table 2** Comparison between optimization algorithms

| Algorithm | Accuracy (cost value) | Time complexity (s) |
|---|---|---|
| Grasshopper [30] | 45.244 | 320 |
| Dragonfly [29] | 47.540 | 425 |
| Bat [39] | 47.552 | 5393 |
| Artificial bee colony [40] | 575.707 | 2182 |
| Simulated annealing [41] | 594.730 | 6315 |
| Harmony [42] | 596.409 | 614 |
| Imperialist [43] | 597.438 | 812 |
| Bees [44] | 602.399 | 1684 |
| Firefly [45] | 605.270 | 1856 |
| Particle swarm optimization [46] | 650.464 | 2064 |
| Differential [47] | 674.832 | 776 |
| Cultural [48] | 926.722 | 1619 |
| Weed [49] | 6.07E+ | 447 |
| Grasshopper [30] | 45.244 | 320 |
| Dragonfly [29] | 47.540 | 425 |
| Bat [39] | 47.552 | 5393 |

classifiers trained using features selected by GOA are denoted as GOA-KNN and GOA-SVM. The following subsections provide details regarding the steps involved in developing these algorithms.

## *3.1 Preprocessing*

Illumination and pose normalization techniques are used in the preprocessing stage to mitigate their negative effects on the overall performance of the algorithm. The normalization technique divides the face image into four sub-segments which are each processed independently. The location of the nose is considered the middle point of the image where this image splitting occurs. Illumination normalization is performed for each segment based on the probability density function of its pixels' grey levels. Upon completing the normalization process, the sub segments are merged and subjected to pixel averaging followed by the application of filters. Details of the entire process are available in [34].

### 3.2 Feature Extraction

Conventional LBP is typically computed for each pixel $(x_c, y_c)$ of an image with the consideration of small circular neighborhood values (with radius $R$ pixels). Let $g_c$ denotes the gray level value of that pixel, then $LBP_{P,R}\ (x_c, y_c)$ is defined as follows

$$LBP_{(P,R)}(x_c, y_c) = \sum_{p=0}^{p=P-1} S\big(g_p - g_c\big)2^P, \tag{19}$$

$$s(g) = \left\{ \begin{matrix} 1, & if g \geq 0 \\ 0, & otherwise \end{matrix} \right\}, \tag{20}$$

where $P$ corresponds to the number of pixels in the neighborhood with radius $R$. A subset of these $2^P$ binary patterns known as uniform patterns have at most two transitions from 0 to 1 (or vice versa). These uniform patterns play an important role in improving recognition. Thus, the total number of output labels generated by mapping patterns of $p$ bits is $p(p-1)+3$. We can mathematically define the uniform LBP $(LBP(P, R)^{u2}$ as:

$$LBP^{u2}_{P,R}(x_c, y_c) = \left\{ \begin{matrix} I\big(LBP_{P,R}(x_c, y_c)\big), if U\big(LBP_{P,R}\big) \leq 2] \\ P(P-1)+2, otherwise \end{matrix} \right., \tag{21}$$

where $I(z) \in [0, P(P-1)+1]$ and

$$U\big(LBP_{P,R}\big) = S(g_{P-1} - g_c) - S(g_0 - g_c) + \sum_{1}^{P} |S(g_P - g_c) - S(g_{P-1} - g_c)| \tag{22}$$

$U\big(LBP_{P,R}\big)$ denotes the pattern's number of spatial bitwise transitions (1/0 changes). If the value of $U\big(LBP_{P,R}\big) < 2$, the corresponding pixel is labeled by an index function *I(Z)*. Otherwise, the pixel will be assigned a value of $(P-1)P+2$. Each uniform pattern is assigned an index based on the index function $I(Z)$ which contains $(P-1)P+2$ indices [35]. The global high-dimensional feature descriptor is then generated by concatenating all the features.

### 3.3 Feature Selection and Classification

Extracting features using ULBP is sensitive to noise and can lead to irrelevant features. The feature extraction method results in a high dimensional feature vector which affects the accuracy and computational cost of a classifier. An efficient FR method could be built by identifying the most important features of the face image.

These problems are solved via feature selection which we will perform using BDA and GOA (presented previously in sections A and B, respectively). The parameters used for BDA and GOA are summarized below:

- BDA
  - Test Size = 1
  - Maximum Iterations = 50
  - Number of Particles = 5
- GOA
  - Maximum Number of Generations = 50
  - Number of Search Agents = 5
  - Lower Bound = −10
  - Upper Bound = 10

The candidate population (number of particles/search agents) for each optimization algorithm is first initialized, then the search for the best features is performed. After each iteration, features which have been identified will be used as inputs to the KNN or SVM classifiers. The resulting recognition accuracy will be used as the fitness function to compare the new set of features to the previous one. Features that lead to the highest accuracy will be selected for facial recognition purposes. We use each optimization algorithm separately alongside each classification algorithm to identify the combination that maximizes recognition accuracy. Feature selection based on the four combinations, BDA-KNN, BDA-SVM, GOA-KNN and GOA-SVM follow similar steps as shown in Algorithm 3.

**Algorithm 3** Feature Selection Process

```
1. Assign a class label to each individual in the da-
   taset, P1, P2,….., Pn
2. Initialize BDA/GOA parameters
3. Train the KNN/SVM algorithm and compute its accuracy
4. Execute Algorithm 1(BDA)/2(GOA)
5. Repeat Step 3 until stopping conditions are satisfied
   or max iterations
```

## 4 Results and Discussion

All experiments described in this section are performed using the Windows 10 on an Intel Core-i5 CPU with 2 GB RAM and MATLAB version 2018. We use three datasets for comparative purposes, the first of which being the Olivetti-Oracle Research Lab

(ORL) face database. The database contains 400 frontal faces, each with a size of 112 X 92 pixels. They can be subdivided into 10 tightly cropped images of 40 individuals with variations in pose, illumination, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The second dataset used is the AR face database created by Aleix Martinez and Robert Benavente. The AR database consists of 4000 color images of 126 different people, divided into 46 females and 70 males. The facial images were taken under restricted conditions but with variations in illumination, facial expression and occlusion with sunglasses, scarves, and hair styles. Labeled Faces in the Wild (LFW) is the third and final dataset used in this work [36]. It consists of 5749 different subjects where 1680 subjects have two or more images, resulting in a total of 13,233 images. Similar to the previously discussed datasets, the images have differences in terms of pose, lighting, expression, background, race, age, gender, clothing, occlusions, camera, focus, and other parameters. This database is considered one of the most vital datasets to analyze the robustness of FR against uncontrolled conditions.

We evaluate the performance of the four combinations, BDA-KNN, BDA-SVM, GOA-KNN and GOA-SVM in terms of their time complexity and accuracy. We first compare the optimized algorithms with their unoptimized counterparts to show the performance gains in terms of both metrics. As seen in Tables 3 and 4, the optimized algorithms displayed significant performance improvements. Reduction of the features from feature selection leads to improved accuracy (by preventing overfitting) and improved time complexity.

Prior to using the reduced feature set, SVM is generally more accurate than KNN albeit being slower. The performance gap between both algorithms is reduced by applying the optimization algorithms for feature selection. In addition, BDA-KNN and GOA-KNN now slightly outperforms BDA-SVM and GOA-SVM respectively. The experiments also indicate that the GOA variants of both classifiers slightly

**Table 3** Performance Improvements of KNN

| Algorithm | KNN | | BDA-KNN | | GOA-KNN | |
|---|---|---|---|---|---|---|
| Metric | T (s) | Acc (%) | T (s) | Acc (%) | T (s) | Acc (%) |
| AR | 5.70 | 98.23 | 3.40 | 99.40 | 3.20 | 99.90 |
| ORL | 3.72 | 98.49 | 2.44 | 99.46 | 2.15 | 99.60 |
| LFW | 5.66 | 98.18 | 3.60 | 99.39 | 3.11 | 99.80 |

**Table 4** Performance Improvements of SVM

| Algorithm | SVM | | BDA-SVM | | GOA-SVM | |
|---|---|---|---|---|---|---|
| Metric | T (s) | Acc (%) | T (s) | Acc (%) | T (s) | Acc (%) |
| AR | 5.70 | 98.23 | 3.40 | 99.40 | 3.20 | 99.90 |
| ORL | 3.72 | 98.49 | 2.44 | 99.46 | 2.15 | 99.60 |
| LFW | 5.66 | 98.18 | 3.60 | 99.39 | 3.11 | 99.80 |

**Table 5** Accuracy (%) comparison with existing work

| Method | Dataset | | | Feature Selection | Classifiers |
|---|---|---|---|---|---|
| | ORL | AR | LFW | Algorithm | |
| GOA-KNN | 99.96 | 99.94 | 99.89 | GOA | KNN |
| BDA-KNN | 99.90 | 99.85 | 99.86 | BDA | KNN |
| GOA-SVM | 99.60 | 99.90 | 99.80 | GOA | SVM |
| BDA-SVM | 99.46 | 99.40 | 99.39 | BDA | SVM |
| [37] | 99.8 | – | – | Genetic | SVM |
| [50] | 95.43 | 98.8 | 83.37 | – | SVM |
| [21] | 98.75 | – | – | PSO | KNN |
| [35] | 98.4 | 98.33 | – | – | KNN |
| [18] | – | 98.32 | – | – | Component classifiers |
| [8] | 97.7 | 92.40 | 60.50 | Firefly | Standard deviation method |
| [12] | 96.5 | – | – | Bat | PCA |
| [23] | 96.5 | – | – | GWO | KNN |
| [38] | 91.5 | – | – | Differential, Genetic | SVM |
| [24] | 97.14 | – | – | Ant colony | Distance measure classifiers |
| [26] | 98 | – | – | PSO | KNN |
| [27] | 96.5 | – | – | Cuckoo search | PCA |

outperform their BDA counterparts. One explanation for this phenomenon is that GOA is more suited to identify global optima whereas BDA tends to generate locally optimal results. We also compare the proposed work against other recently proposed approaches based on recognition rate as shown in Table 5. For all datasets, BDA-KNN, BDA-SVM, GOA-KNN and GOA-SVM generally outperform their peers.

The new optimization algorithms were effective in removing irrelevant, noisy, and redundant features that were extracted using ULBP. This is apparent from the high prediction accuracy of the proposed method as compared to other FR proposals in Table 5. This result also supports our findings in Table 2, which identified that the dragonfly and grasshopper algorithms outperform other optimization algorithms. To the best of our knowledge, the proposed work is one of the first in investigating the use of both dragonfly and grasshopper algorithms specifically for facial recognition purposes.

## 5 Conclusion

In this paper, we investigate the application of two relatively new optimization algorithms in facial recognition, the dragonfly and grasshopper optimization algorithms.

We select these algorithms after performing a thorough comparison with 13 of its peers in terms of feature selection capability. Both algorithms are then used for feature selection alongside two classifiers, k-nearest neighbor and support vector machine. We denote the combination of these approaches as BDA-KNN, BDA-SVM, GOA-KNN and GOA-SVM respectively. As expected, significant performance improvements were obtained when the optimized algorithms were compared to their unoptimized counterparts. Interestingly, we also found that the KNN outperformed their SVM counterparts after application of the optimization algorithms for feature selection, whereas the inverse held true prior to feature selection. We also found that the GOA-based classifiers outperform their BDA counterparts due to the capability of GOA in identifying globally optimal solutions as compared to the locally optimal solutions generated by BDA. Performance comparison against other similar approaches in literature depicts the superiority of the proposed methods in terms of both accuracy and time complexity. Moving forward, our findings imply that future facial recognition algorithms should leverage upon grasshopper optimization for feature selection to maximize performance.

**Acknowledgements** This work is supported in part by the Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS), project number FRGS/1/2019/ICT05/USM/02/1 and Universiti Sains Malaysia under grant no. 8011036.

## References

1. Ma H, Celik T (2019) FER-Net: facial expression recognition using densely connected convolutional network. Electron Lett 55(4):184–186
2. Chen Z, Huang W, Lv Z (2017) Towards a face recognition method based on uncorrelated discriminant sparse preserving projection. Multimed Tools Appl 76(17):17669–17683
3. Chengeta K, Viriri S (2018) A survey on facial recognition based on local directional and local binary patterns. In: 2018 conference on information communications technology and society (ICTAS), pp 1–6. IEEE
4. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
5. Yan Z, Yuan C (2004) Ant colony optimization for feature selection in face recognition. In: International conference on biometric authentication, pp 221–226. Springer, Heidelberg
6. Connolly JF, Granger E, Sabourin R (2012) Evolution of heterogeneous ensembles through dynamic particle swarm optimization for video-based face recognition. Pattern Recogn 45(7):2460–2477
7. Jakhar R, Kaur N, Singh R (2011) Face recognition using bacteria foraging optimization-based selected features. Int J Adv Comput Sci Appl 1(3)
8. Agarwal V, Bhanot S (2018) Radial basis function neural network-based face recognition using firefly algorithm. Neural Comput Appl 30(8):2643–2660
9. Islam KT, Raj RG, Al-Murad A (2017) Performance of SVM, CNN, and ANN with BoW, HOG, and image pixels in face recognition. In: 2017 2nd international conference on electrical & electronic engineering (ICEEE), pp 1–4. IEEE
10. Kumar M, Jindal MK, Sharma RK (2011) k-nearest neighbor based offline handwritten Gurmukhi character recognition. In: 2011 international conference on image information processing, pp 1–4. IEEE

11. Parveen P, Thuraisingham B (2006) Face recognition using multiple classifiers. In: 2006 18th IEEE international conference on tools with artificial intelligence (ICTAI 2006), pp 179–186. IEEE
12. Kumar D (2017) Feature selection for face recognition using DCT-PCA and Bat algorithm. Int J Inf Technol 9(4):411–423
13. Sinha P, Sinha P (2015) Comparative study of chronic kidney disease prediction using KNN and SVM. Int J Eng Res Technol 4(12):608–612
14. Phillips PJ (1999) Support vector machines applied to face recognition. In: Advances in neural information processing systems, pp 803–809
15. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):1–27
16. Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. Multimed Tools Appl 76(6):7803–7821
17. Gao Y, Lee HJ (2019) Pose-invariant features and personalized correspondence learning for face recognition. Neural Comput Appl 31(1):607–616
18. Zhu Y, Xue J (2017) Face recognition based on random subspace method and tensor subspace analysis. Neural Comput Appl 28(2):233–244
19. Lu Y, Cui J, Fang X (2014) Enhancing sparsity via full rank decomposition for robust face recognition. Neural Comput Appl 25(5):1043–1052
20. Shukla AK, Kanungo S (2019) An automated face retrieval system using grasshopper optimization algorithm-based feature selection method. In: International conference on emerging current trends in computing and expert technology, pp 492–502. Springer, Cham
21. Sasirekha K, Thangavel K (2019) Optimization of K-nearest neighbor using particle swarm optimization for face recognition. Neural Comput Appl 31(11):7935–7944
22. Gupta A, Goel L (2016) Heuristic approach for face recognition using artificial bee colony optimization. In: The international symposium on intelligent systems technologies and applications, pp 209–223. Springer, Cham
23. Abd AL, El-Hafeez T, Zaki AM (2018) Face recognition based on Grey Wolf optimization for feature selection. International conference on advanced intelligent systems and informatics. Springer, Cham, pp 273–283
24. Aro T, Abikoye O, Oladipo I, Awotunde B (2019) Enhanced Gabor features based facial recognition using ant colony optimization algorithm. J Sustain Technol 10(1)
25. Benamara NK, Zigh E, Stambouli TB, Keche M (2019) Efficient Multispectral face recognition using random feature selection and PSO-SVM. In: Proceedings of the 2nd international conference on networking, information systems & security, pp 1–6
26. Eleyan A (2019) Particle swarm optimization based feature selection for face recognition. In: 2019 seventh international conference on digital information processing and communications (ICDIPC), pp 1–4. IEEE
27. Malhotra P, Kumar D (2019) An optimized face recognition system using cuckoo search. J Intell Syst 28(2):321–332
28. Král P, Vrba A, Lenc L (2019) Enhanced local binary patterns for automatic face recognition. In: International conference on artificial intelligence and soft computing, pp 27–36. Springer, Cham
29. Mirjalili S (2016) Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Comput Appl 27(4):1053–1073
30. Saremi S, Mirjalili S, Lewis A (2017) Grasshopper optimisation algorithm: theory and application. Adv Eng Softw 105:30–47
31. Mirjalili SZ, Mirjalili S, Saremi S, Faris H, Aljarah I (2018) Grasshopper optimization algorithm for multi-objective optimization problems. Appl Intell 48(4):805–820
32. Neve AG, Kakandikar GM, Kulkarni O (2017) Application of grasshopper optimization algorithm for constrained and unconstrained test functions. Int J Swarm Intell Evol Comput 6(165):2
33. Virtual library of simulation experiments: test functions and datasets

34. Sharif M, Mohsin S, Jamal MJ, Raza M (2010) Illumination normalization preprocessing for face recognition. In: 2010 the 2nd conference on environmental science and information application technology, vol 2, pp 44–47. IEEE
35. Salyut J, Kurnaz C (2018) Profile face recognition using local binary patterns with artificial neural network. In: 2018 international conference on artificial intelligence and data processing (IDAP), pp 1–4. IEEE
36. Learned-Miller E, Huang GB, Roy Chowdhury A, Li H, Hua G (2016) Labeled faces in the wild: a survey. In: Advances in face detection and facial image analysis, pp 189–248. Springer, Cham
37. Singh G, Chhabra I (2018) Genetic algorithm implementation to optimize the hybridization of feature extraction and metaheuristic classifiers. In: Hybrid metaheuristics for image analysis, pp 49–86. Springer, Cham
38. Maheshwari R, Kumar M, Kumar S (2016) Optimization of feature selection in face recognition system using differential evolution and genetic algorithm. In: Proceedings of fifth international conference on soft computing for problem solving, pp 363–374. Springer, Singapore
39. Yang XS (2010) A new metaheuristic bat-inspired algorithm. In: Nature inspired cooperative strategies for optimization (NICSO 2010), pp 65–74. Springer, Heidelberg
40. Kiran MS (2014) Improved artificial bee colony algorithm for continuous optimization problems. J Comput Commun 2(04):108
41. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
42. Yang XS (2009) Harmony search as a metaheuristic algorithm. In: Music-inspired harmony search algorithm, pp 1–14. Springer, Heidelberg
43. Atashpaz-Gargari E, Lucas C (2007) Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. In: 2007 IEEE congress on evolutionary computation, pp 4661–4667. IEEE
44. Pham DT, Castellani M (2015) A comparative study of the Bees Algorithm as a tool for function optimisation. Cogent Eng 2(1):1091540
45. Yang XS (2009) Firefly algorithms for multimodal optimization. In: International symposium on stochastic algorithms, pp 169–178. Springer, Heidelberg
46. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks, vol 4, pp 1942–1948. IEEE
47. Storn R, Price K (1997) Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359
48. dos Reis Ribeiro M, de Aguiar MS (2011) Cultural Algorithms: a study of concepts and approaches. In: 2011 workshop-school on theoretical computer science, pp 145–148. IEEE
49. Mehrabian AR, Lucas C (2006) A novel numerical optimization algorithm inspired from weed colonization. Ecol Inform 1(4):355–366
50. Vinay A, Shekhar VS, Manjunath N, Murthy KB, Natarajan S (2018) Expediting automated face recognition using the novel ORB 2-IPR framework. In: Proceedings of international conference on cognition and recognition, pp 223–232. Springer, Singapore

# Optimization on the Financial Management of Banks with Two-Stage Goal Programming Model


**Lam Weng Siew, Lam Weng Hoe, and Chen Jia Wai**

**Abstract** The strategic planning is important in bank financial management. The banks and financial institutions have to achieve multiple goals in optimizing asset, liability, equity, earnings, profit and optimum management item. The subjective judgment in assigning weight of the goals is a drawback in financial management of the banks since it may cause inconsistent judgment. In addition, there are no comprehensive studies done on comparison among the banks for benchmarking based on the past studies in optimizing the financial management. Hence, this paper aims to improve the bank financial management by proposing a two-stage goal programming (GP) model to compare and optimize the bank financial management. The proposed model is developed based on entropy method in determining the weight of the goal at the first stage before optimizing the financial management with GP model at the second stage. Four listed banks in Malaysia are investigated in this study. The results indicate that the goal for asset, equity and optimum management item have been achieved by all banks. Furthermore, the target value of asset, equity, earning and profit can be increased according to the optimal solution of the proposed model. The significance of this paper is to provide insights to the banks for further improvement based on the optimal solution of the proposed model.

**Keywords** Entropy · Goal programming · Optimization · Financial management · Potential improvement

L. W. Siew (✉) · L. W. Hoe · C. J. Wai
Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia
e-mail: lamws@utar.edu.my

L. W. Hoe
e-mail: whlam@utar.edu.my

C. J. Wai
e-mail: jiawai_chen@hotmail.com

L. W. Siew · L. W. Hoe
Centre for Business and Management, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_11

## 1 Introduction

Banks play a significant role in financial market as well as economies in every country [1]. Monitoring on the bank financial management is necessary to meet the target profit as well as control the liquidity. Thus, data analysis on financial statement is essential in financial planning [2]. Bank financial management considers multiple goals such as asset, liability, equity, earnings, profitability and optimum management item [3]. A trade-off point must be obtained among multiple goals in order to optimize the financial management [4]. Decision makers have to achieve an optimal solution that best fit their desire goals. Therefore, goal programming (GP) model has been presented to solve multiple goals in optimizing the financial management of banks [5, 6].

In bank financial management, there are some limitations and drawbacks identified based on the past studies. Firstly, there is no comprehensive study done on comparison among the banks for benchmarking using GP model. GP model is only constructed to optimize the financial management of a single bank without benchmarking on other banks. The comparison among the banks in financial management is important since it can determine the potential improvement according to the benchmark [7–9]. Besides that, subjective judgment in assigning weight of the goals is another drawback in optimizing the financial management of banks. The problem of reliability on respondents' judgment or subjective weights may result in bias perspective and inconsistent issue [10]. Hence, this study aims to propose a two-stage GP model in optimizing the bank financial management. At the first stage, entropy method is proposed to determine the objective weight of each goal. At the second stage, a GP model is constructed to optimize the bank financial management with multiple goals based on the entropy weights and benchmark target value.

## 2 Literature Review

Kosmidou and Zopounidis [2] developed a GP model to examine the asset and liability of a Greek commercial bank. The authors utilized the financial statement in 1999 to develop an optimal asset and liability management for the following year. Several contradictory goals such as liquidity and returns were considered in the model. Different goals and constraints were taken into account as well as deviation variables to determine the favorable scenarios for future. In short, the proposed model aims to seek for a direction for bank's future financial planning.

Naderi [3] denoted that company's future risk can be predetermined with GP model. The optimal asset and liability management can be achieved through proper structure of bank's balance sheet elements. The findings proved that the model managed to determine appropriate structure for the optimal management.

Halim et al. [5] developed a GP model to examine the achievement and improvement of six goals in bank financial management. Data for the main goals from 2010 to

2014 were obtained through annual statement. The results showed that all goals were achievable and four goals could be increased to higher aspiration level. Therefore, potential improvements on goal achievement can be identified with GP model.

Tektas [11] applied GP model in determining an efficient financial management for the bank with respect to different managerial strategies. Two Turkish commercial banks were investigated to improve the wellbeing of the banks. Different goals such as liquidity, asset and revenue were examined to determine the optimal financial management of the banks.

Arewa et al. [12] conducted a research in United Bank of Africa to investigate the financial management of the bank using GP model. Six goals such as assets, liabilities, equities, profits, earnings and optimum management item were examined to determine the deviation of each goal in optimizing the proportions of items in financial statement.

Viswanathan et al. [13] indicated that GP model is a useful tool for asset allocation and liability composition because it considers large amount of constraints to determine the optimal solution for the bank. The generated results managed to show a realistic and compatible composition for asset and liability.

Chen et al. [14] developed a GP model to analyze the financial management of a listed bank for the period of 2011–2015. Six goals such as assets, liabilities, equities, profits, earnings and optimum management item were examined in their studies. The results showed that all goals were achievable and three goals could be increased to new target value for further improvement.

In addition, GP model has also been utilized to solve multiple objective decision problem in portfolio optimization [15–18].

## 3 Data and Methodology

### *3.1 Data*

This study investigates four listed banks in Malaysia, namely Public Bank Berhad (PBBANK), RHB Bank Berhad (RHBBANK), CIMB Group Holding Berhad (CIMB) and Malayan Banking Berhad (MAYBANK) from year 2012 to 2016. Asset, liability, equity, earning, profitability and optimum management item are the main goals in financial management of the banks [3, 5].

### *3.2 Proposed Two-Stage Goal Programming Model*

A two-stage GP model is proposed to compare and optimize multiple goals in bank financial management. The proposed two-stage GP model comprises two stages as follows.

Stage 1:
At the first stage, the weight of the goal is determined with entropy approach.
Stage 2:
At the second stage, a GP model is constructed for each bank to optimize multiple goals based on the entropy weights obtained in the first stage. The potential improvement will be recommended based on the optimal solution obtained.

**Entropy Method (First Stage)**
Objective weight is emphasized owing to subjective weighting from the decision makers are based on their opinions and preferences. Therefore, the subjective judgment might be imprecise or inconsistent. Entropy method determines the objective weight of each goal [19, 20]. From the past studies, entropy weight has been integrated with TOPSIS model in multiple-criteria decision making (MCDM) problems [1, 4, 10, 19–21]. In this study, entropy method is proposed to determine the weights of the goals as follows.

Step 1: Form the decision matrix based on the total performance rating of all banks under each criterion. The rows indicate alternatives for $i = 1, 2, 3, \ldots, n$ whereas columns refer to criteria for $j = 1, 2, 3, \ldots, m$.

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix}$$

Step 2: Form a new normalized matrix ($Z$) by normalizing the decision matrix based on Eq. (1).

$$z_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} y_{ij}} \tag{1}$$

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

Step 3: Determine the entropy value ($e_j$).

$$e_j = -k \sum_{i=1}^{n} z_{ij} \ln z_{ij} \text{ where } k = \frac{1}{\ln n} \tag{2}$$

A higher entropy value proposes less useful information about the criteria, hence a smaller weight will be assigned and vice versa [20].

Step 4: Obtain the degree of divergence ($d_j$) of the intrinsic information involved by each criterion.

$$d_j = 1 - e_j \tag{3}$$

Step 5: Determine the objective weight of each criterion.

$$w_j = \frac{d_j}{\sum_{k=1}^{m} d_k} \tag{4}$$

**Goal Programming Model (Second Stage)**

At the second stage, a GP model is constructed for each bank to optimize multiple goals based on the entropy weights and benchmark target value. GP model is designed owing to the difficulty in solving multiple objectives especially in achieving contradict goals for a company. GP model is able to deal with multiple goals simultaneously while obtaining the optimal solution that satisfied all the restrictions on the decision variables [17, 22, 23]. The GP model is shown as follows:

$$\text{Minimize } z = w_1 G_1 + w_2 G_2 + \cdots + w_j G_j \text{ where } j = 1, 2, 3, \ldots, m. \tag{5}$$

Subject to

$$\sum_{c=1}^{h} (a_{jc} x_c + d_j^- - d_j^+) = g_j \tag{6}$$

$$x_c, d_j^-, d_j^+ \geq 0$$

where

$z$ total deviation;
$w_j$ entropy weights for $j = 1, 2, 3, \ldots, m$;
$d_j^-$ negative deviation variable (underachievement) for $j = 1, 2, 3, \ldots, m$;
$d_j^+$ positive deviation variable (overachievement) for $j = 1, 2, 3, \ldots, m$;
$x_c$ amount of financial statement for $c = 1, 2, 3, \ldots, h$;
$a_{jc}$ weight of financial statement for $c = 1, 2, 3, \ldots, h$;
$g_j$ target value for $j = 1, 2, 3, \ldots, m$.

This study aims to optimize multiple goals as follows.

$G1$ Maximize asset ($d_1^-$),
$G2$ Minimize liability ($d_2^+$),
$G3$ Maximize equity ($d_3^-$),
$G4$ Maximize profitability ($d_4^-$),
$G5$ Maximize earnings ($d_5^-$),

*G*6 Maximize optimum management item ($d_6^-$).

LINGO software is employed to solve the proposed two-stage GP model. According to the optimal solution of GP model, the goal has been achieved if the respective deviation variable is zero [5]. The potential improvement can be determined for each bank according to the deviation from the target value. LINGO software has been applied in various studies to solve the optimization problem [24–31].

## 4 Results

Table 1 presents the weight of multiple goals as described in the first stage of the proposed model.

As shown in Table 1, the highest weight of the goal is earnings (0.1973), followed by profit (0.1949), equity (0.1587), optimum management item (0.1503), asset (0.1497) and lastly liability (0.1490). Based on the Entropy method, the decision makers can identify the importance of the goals based on past financial data as data driven decision analysis instead of obtaining subjective judgment.

Tables 2, 3, 4 and 5 present the optimal value and goal achievement for CIMB, MAYBANK, PBBANK and RHBBANK respectively according to the optimal solution of GP model.

**Table 1** Weight of multiple goals

| | Asset | Liability | Equity | Profit | Earnings | Optimum management item |
|---|---|---|---|---|---|---|
| Weight | 0.1497 | 0.1490 | 0.1587 | 0.1949 | 0.1973 | 0.1503 |

**Table 2** Optimal solution of CIMB (RM Trillion)

| Goals | Target value | Optimal value | $d_j^-$ | $d_j^+$ | Goal status |
|---|---|---|---|---|---|
| Maximize asset | 3.1399 | 3.1399 | 0 | 0 | Achieved |
| Minimize liability | 0.9709 | 2.8383 | 0 | 1.8674 | Not achieved |
| Maximize equity | 0.2804 | 0.3016 | 0 | $2.1215 \times 10^{-2}$ | Achieved |
| Maximize profit | 0.0335 | 0.0249 | $8.5566 \times 10^{-3}$ | 0 | Not achieved |
| Maximize earnings | 0.0972 | 0.1058 | 0 | $8.5584 \times 10^{-3}$ | Achieved |
| Maximize optimum management item | 6.4105 | 6.4105 | 0 | 0 | Achieved |

**Table 3** Optimal solution of MAYBANK (RM Trillion)

| Goals | Target value | Optimal value | $d_j^-$ | $d_j^+$ | Goal status |
|---|---|---|---|---|---|
| Maximize asset | 3.1399 | 3.1410 | 0 | $1.0974 \times 10^{-3}$ | Achieved |
| Minimize liability | 0.9709 | 2.8448 | 0 | 1.8739 | Not achieved |
| Maximize equity | 0.2804 | 0.2962 | 0 | $1.5822 \times 10^{-2}$ | Achieved |
| Maximize profit | 0.0335 | 0.0313 | $2.1985 \times 10^{-3}$ | 0 | Not achieved |
| Maximize earnings | 0.0972 | 0.0972 | 0 | 0 | Achieved |
| Maximize optimum management item | 6.4105 | 6.4105 | 0 | 0 | Achieved |

**Table 4** Optimal solution of PBBANK (RM Trillion)

| Goals | Target value | Optimal value | $d_j^-$ | $d_j^+$ | Goal status |
|---|---|---|---|---|---|
| Maximize asset | 3.1399 | 3.1423 | 0 | $2.4248 \times 10^{-3}$ | Achieved |
| Minimize liability | 0.9709 | 2.8500 | 0 | 1.8791 | Not achieved |
| Maximize equity | 0.2804 | 0.2924 | 0 | $1.1969 \times 10^{-3}$ | Achieved |
| Maximize profit | 0.0335 | 0.0435 | 0 | $9.9983 \times 10^{-3}$ | Achieved |
| Maximize earnings | 0.0972 | 0.0823 | $1.4855 \times 10^{-2}$ | 0 | Not achieved |
| Maximize optimum management item | 6.4105 | 6.4105 | 0 | 0 | Achieved |

According to Table 2, CIMB is able to achieve four goals for asset, equity, earnings and optimum management item because of zero values for negative deviation respectively. This implies that CIMB is able to meet or overachieve the target values for these goals. For liability and profitability goals, $d_2^+$ and $d_4^-$ are non-zero. Therefore, these two goals are not achieved for CIMB. From the findings, the target value of equity and earnings can be revised for continuous improvement. The target value of equity and earnings can be improved by RM 0.0212 trillion and RM 0.0086 trillion respectively based on the positive deviation. As shown in Table 3, the goal achievements for MAYBANK are same as CIMB, which are asset, equity, earnings and

**Table 5** Optimal solution of RHBBANK (RM Trillion)

| Goals | Target value | Optimal value | $d_j^-$ | $d_j^+$ | Goal status |
|---|---|---|---|---|---|
| Maximize asset | 3.1399 | 3.1418 | 0 | $1.8838 \times 10^{-3}$ | Achieved |
| Minimize liability | 0.9709 | 2.8614 | 0 | 1.8905 | Not achieved |
| Maximize equity | 0.2804 | 0.2804 | 0 | 0 | Achieved |
| Maximize profit | 0.0335 | 0.0298 | $3.7242 \times 10^{-3}$ | 0 | Not achieved |
| Maximize earnings | 0.0972 | 0.0972 | 0 | 0 | Achieved |
| Maximize optimum management item | 6.4105 | 6.4105 | 0 | 0 | Achieved |

optimum management item. The target value of asset and equity can be improved by RM 0.0011 trillion and RM 0.0158 trillion respectively for MAYBANK in future.

From the results in Table 4, PBBANK is able to achieve four goals as well except liability and earnings. In addition, the target value of asset, equity and profit can be improved by RM 0.0024 trillion, RM 0.0120 trillion and RM 0.0100 trillion respectively over the next five years for continuous improvement. For RHBBANK, $d_1^-, d_3^-, d_5^-$ and $d_6^-$ are zero as presented in Table 5. This implies that the asset, equity, earnings and optimum management item are achieved by RHBBANK. Moreover, the target value of asset can be improved by $1.8838 \times 10^{-3}$ based on the optimal solution.

Table 6 shows the potential improvement for each bank according to the deviation from the benchmark target value.

As shown in Table 6, the potential improvement can be determined in order to achieve the benchmark target value. Three goals namely, asset, equity and optimum management item are achieved by all banks. However, liability goal is unachievable

**Table 6** Potential improvement according to the deviation from the benchmark target value

| Goals | CIMB (RM trillion) | MAYBANK (RM trillion) | PBBANK (RM trillion) | RHBBANK (RM trillion) |
|---|---|---|---|---|
| Maximize asset | 0 | 0 | 0 | 0 |
| Minimize liability | 1.8674 | 1.8739 | 1.8791 | 1.8905 |
| Maximize equity | 0 | 0 | 0 | 0 |
| Maximize profit | $8.5566 \times 10^{-3}$ | $2.1985 \times 10^{-3}$ | 0 | $3.7242 \times 10^{-3}$ |
| Maximize earnings | 0 | 0 | $1.4855 \times 10^{-2}$ | 0 |
| Maximize optimum management item | 0 | 0 | 0 | 0 |

by all banks because the positive deviation from the benchmark target value is non-zero. CIMB, MAYBANK, PBBANK and RHBBANK should minimize their total liability by RM 1.8674 trillion, RM 1.8739 trillion, RM 1.8791 trillion and RM 1.8905 trillion respectively to meet the benchmark target value of RM 0.9709 trillion. For profitability, only PBBANK can achieve the goal due to zero deviation from the benchmark target value. CIMB, MAYBANK and RHBBANK should decrease RM 0.0086 trillion, RM 0.0022 trillion and RM 0.0037 trillion respectively to meet the benchmark target value of RM 0.0335 trillion.

## 5 Conclusion

A two-stage GP model is proposed to compare and optimize multiple goals in financial management of the banks in Malaysia. The weight of each goal is determined using entropy method at the first stage. The findings indicate that CIMB, MAYBANK, PBBANK and RHBBANK can achieve the goal for asset, equity and optimum management item. Furthermore, the target value of asset, equity, earning and profit can be increased according to the optimal solution of the proposed model. The significance of this study is to determine the potential improvement on liability, profit as well as earnings in bank financial management.

## References

1. Elsayed E, Dawood A, Karthikeyan R (2017) Evaluating alternatives through the application of topsis method with entropy weight. Int J Eng Trends Technol 46(2):60–66
2. Kosmidou K, Zopounidis C (2002) A multi objective methodology for bank asset liability management, financial engineering, e-commerce and supply chain. Kluwer Academic Publishers, Dordrecht
3. Naderi S, Minouei M, Gashti H (2013) Asset and liability optimal management mathematical modeling for bank. J Basic Appl Sci Res 3(1):484–493
4. Elsayed E, Dawood A, Karthikeyan R (2017) Using Vikor technique for evaluating customer satisfaction in bank using entropy weight. Int J Innovative Res Sci Eng Technol 6(5):9655–9663
5. Halim B, Karim H, Fahami N, Mahad N, Nordin S, Hassan N (2015) Bank financial statement management using a goal programming model. Procedia Soc Behav Sci 211:498–504
6. Zaloom V, Tolga A, Chu H (1986) Bank funds management by goal programming. Comput Ind Eng 11(1–4):132–135
7. Feroz EH, Kim S, Raab RL (2003) Financial statement analysis: a data envelopment analysis approach. J Oper Res Soc 54(1):48–58
8. Liew KF, Lam WS, Lam WH (2017) An empirical evaluation on the efficiency of the companies in Malaysia with data envelopment analysis model. Adv Sci Lett 23(9):8264–8267
9. Tehrani R, Mehragan MR, Golkani MR (2012) A model for evaluating financial performance of companies by data envelopment analysis. Int Bus Res 5(8):8–16
10. Deng H, Yeh C, Willis R (2000) Inter-company comparison using modified TOPSIS with objective weights. Comput Oper Res 27:963–973
11. Tektas A, Ozkan-Gunay EN, Gunay G (2005) Asset and liability management in financial crisis. J Risk Finance 6(2):135–149

12. Arewa A, Owoputi J, Torbira L (2013) Financial statement management, liability reduction and asset accumulation: an application of goal programming model to a Nigerian bank. Int J Financ Res 4(4):83–90
13. Viswanathan PK, Balasubramanian G (2014) Modeling asset allocation and liability composition for Indian banks. Manag Financ 40(7):700–723
14. Chen JW, Lam WS, Lam WH (2019) Mathematical modelling of bank financial management in Malaysia using goal programming approach. In: Proceedings of the third international conference on computing, mathematics and statistics. Springer Nature, Singapore, pp 119–125
15. Jaaman SH, Lam WH, Isa I (2014) A new higher moment portfolio optimisation model with conditional value at risk. Int J Oper Res 21(4):451–465
16. Lam WS, Lam WH (2016) Strategic decision making in portfolio management with goal programming model. Am J Oper Manag Inf Syst 1(1):34–38
17. Lam WS, Jaaman SH, Ismail H (2014) Portfolio optimization in enhanced index tracking with goal programming approach. AIP Conf Proc 1614:968–972
18. Wu LC, Chou SC, Yang CC, Ong CS (2007) Enhanced index investing based on goal programming. J Portfolio Manag 33(3):49–56
19. Huang J (2008) Combining entropy weight and TOPSIS method for information system selection. In: International conference on automation and logistics. Proceedings of the IEEE, Qingdao, China, pp 1965–1968
20. Li X, Gao Z (2015) Application of improved entropy TOPSIS to competitive performance evaluation of power companies. In: International conference on computational science and engineering. Atlantis Press, Paris, pp 183–188
21. Lam WS, Liew KF, Lam WH (2019) Investigation on the performance of construction companies in Malaysia with entropy-TOPSIS model. IOP Conf Ser Earth Environ Sci 385:012006
22. Charnes A, Cooper W, Ferguson R (1955) Optimal estimation of executive compensation by linear programming. Manag Sci 1(2):138–151
23. Winston W (2003) Operations research: applications and algorithms. Cengage Learning, Boston
24. Lam WS, Jaaman SH, Lam WH (2019) Enhanced index tracking with entropy maximization. Adv Appl Stat 53(3):243–258
25. Lam WS, Jaaman SH, Lam WH (2017) Enhanced index tracking in portfolio optimization with two-stage mixed integer programming model. J Fundam Appl Sci 9(5S):1–12
26. Lam WS, Jaaman SH, Lam WH (2019) An enhanced mean-gini extended model in portfolio optimization with different level of risk aversion. ASM Sci J 12(6):41–46
27. Lam WS, Jaaman SH, Lam WH (2020) Portfolio optimization of financial companies with fuzzy TOPSIS-mean-semi absolute deviation model. J Adv Res Dyn Control Syst 12(4S):1488–1495
28. Lam WS, Liew KF, Lam WH (2018) Investigation on the efficiency of financial companies in Malaysia with data envelopment analysis model. J Phys Conf Ser 995:012021
29. Lam WS, Liew KF, Lam WH (2018) An optimal control on the efficiency of technology companies in Malaysia with data envelopment analysis model. J Telecommun Electron Comput Eng 10(1):107–111
30. Lam WS, Jaaman SH, Ismail H (2015) An empirical comparison of different optimization models in enhanced index tracking problem. Adv Sci Lett 21(5):1278–1281
31. Lam WS, Jaaman SH, Ismail H (2015) The impact of human behaviour towards portfolio selection in Malaysia. Procedia Soc Behav Sci 172:674–678

# Evaluating the Performance of Selected Mortality Forecasting Models: A Malaysia Case Study


**Khairunnisa Mokhtar, Syazreen Niza Shair, and Norazliani Md Lazam**

**Abstract** The study of human mortality is growing in Malaysia, as accurate mortality rates are classified important especially for social policy planning. This research aims at evaluating the performance of three selected mortality forecasting models, namely the Lee-Carter, CBD and M8 model in which the two latter models are from Cairns, Blake and Dowd. We applied the Malaysian central death rates and the number of mid-year exposures to the models and estimate the goodness of fits of all models using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In addition, the 20-year out-samples forecast errors are estimated from 1999 to 2019 using the Root Mean Square Errors (RMSEs) and the Mean Absolute Percentage Errors (MAPEs). The findings of this study suggest that the M8 model is the best model for fitting Malaysian mortality data with minimum AIC and BIC values, and by far the most accurate model with the lowest out-sample errors, particularly for higher age category.

**Keywords** The Lee-Carter model · The Cairns, Blake and Dowd model · Mortality modeling

## 1 Introduction

In Malaysia, the study of population mortality is increasing over the past few years as mortality estimates are useful to pension and insurance industries [1]. The importance of the estimation of mortality rates derives from their potential in evaluating the liabilities of pension funding and insurance accurately as well as improving public's

K. Mokhtar · S. N. Shair (✉) · N. M. Lazam
Centre for Actuarial Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
e-mail: syazreen@uitm.edu.my

K. Mokhtar
e-mail: khairunnisamokhtar8@gmail.com

N. M. Lazam
e-mail: norazliani@tmsk.uitm.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_12

health. Statistics have shown that life expectancies in Malaysia have significantly increased from 62.8 in 1966 to 74.5 years old in 2019 [2]. The increase in life expectancy in Malaysia may be resulted from several factors such as better access to health care resources and higher economic status among population [3].

An increase in life expectancy practically gives a good perspective that Malaysian can live longer. However, this also can be directed to another problem known as longevity risk [4]. For example, longevity risk may put pressure to the governments in spending more on health and other social programs especially for elderlies. In terms of individual prospective, an increase in life expectancy provides serious economic consequences as when people get older, they tend to live with chronic health conditions hence will incur substantial costs of treatment. Thus, information on how long one is expected to live is important. To obtain an accurate life expectancy estimate, an accurate mortality forecasting model is much needed particularly for the Malaysian population.

To this date, the Lee-Carter model is known as the significant model which was widely used to forecast mortality rates [5]. This well-developed model is proposed by Ronald D. Lee and Lawrence Carter in 1992 [6]. This model comprises log-bilinear form of mortality, integrating two major effects which are age ($a_x$ and $b_x$) and year ($k_t$) [7, 8]. To reduce the complexity in estimating the two factors ($b_x$ and $k_t$), the model is reduced to a univariate time series forecasting using the Singular Values Decomposition (SVD) technique. Since $k_t$ is the only parameter with time involved, the Autoregressive Integrated Moving Average (ARIMA) models are employed to predict the $k_t$ parameter. The model assumes invariant age-component and a linear time-component reduction over time [6].

The Lee-Carter model not only proven to be well-developed, it has been widely applied in many countries [9, 10] including Malaysia. However, some researchers have proved that the model has some weaknesses such as the age component is assumed invariant over time [11]. An extension version of the Lee-Carter forecasting mortality model denoted as the CBD model has been developed by [12]. This model is unique due to the focus that is more on longevity risk factor. This model contributed a huge influence for financial sectors like retirement funds, life-insurance corporation and independent annuity suppliers [12]. The model accustoms to the development of mortality curve through time with two-factor stochastic model. Instead of one $k_t$, the model is extended to include ${k_t}^{(2)}$. The first $k_t$ affects the dynamics of mortality at all ages in the same way as the Lee-Carter model while the second $k_t$ affects the dynamics of mortality at higher ages [12].

The Lee-Carter model has initiated the development of all the CBD models. The model assumes there is a stationary age function, $a_x$ with no cohort effects and named as the M1 model. Then, from the M1 model, the study from [13] extended the model to include the cohort effect and add one more age related parameter, ${b_x}^{(3)}$. This model was called M2 model. The model M3 is a special case of the model M2 with ${b_x}^{(2)} = 1$ and ${b_x}^{(3)} = 1$ parameter. In order to fit the mortality surface, the M4 model is proposed to include B-splines and P-splines smoothing techniques for age variables. A research from [14] revised the original CBD model to become the M5 model. The M5 model comprises two-period effects (${k_t}^{(1)}$ and ${k_t}^{(2)}$) including age effect and no

component of cohort effect. Next M5 is expanded to the model M6—the first version that integrates cohort effects into the CBD model. The M6 model is extended to the second generation named as the M7 model. This model includes the quadratic term in age factors and the third period effects.

The M8 model is the recent extension of CBD model that incorporates two models which are the M2 and M5 models [15]. The current version of the CBD models are not widely applied by the researchers as they are new. Such models were tested for data from England and Wales and it showed that the M8 model was not only capable of capturing the cohort effect substantially but also had the lowest error [14]. In the comparison study done by utilizing the Italian death rates with regard to Lee-Carter and CBD model, a significant difference can be found between the two. While the CBD model is suitable for older age groups, the Lee-Carter model performs best throughout the whole analysis. These earlier studies show that mortality data from different countries will yield different results where the latest CBD model, the M8 model, is the most accurate model for forecasting mortality rates.

The design of the CBD model based on the risk of longevity suggests that the model might be suitable to apply to Malaysian data as it is predicted by 2030 Malaysia will become an aging nation. Malaysia is currently focusing on social protection arrangements for elderlies in order to prepare the nation with ageing issues [16]. There is a recent study from [17] that made a comparison between the Lee-Carter model and Cairns, Blake and Dows (CBD) model. The study found that, the evaluation between both models showed that there is no model better than the other. It is noteworthy that [17] did not include the most recent extended model which is the M8 model. Hence, this study would like to extend [17] by adding the M8 model and conduct this research using more up-to-date data from 1960 to 2019.

## 2 Methodology

### *2.1 Data*

This section discusses on the selected mortality forecasting models specifically the Lee-Carter model, the CBD model a well as the M8 model. The central mortality rates by age groups for both female and male from 1966 to 2019 (54-year-data) was taken from the Department of Statistics Malaysia (DOSM). Meanwhile the mid-year exposures or number of population by age groups for both genders over the same years was retrieved from the World Population Prospect [18].

## 2.2 The Lee-Carter Model

The Lee-Carter model assumes there are fixed age function and a unique non-parametric age-period. This model is widely used by many researchers since it was proposed. The general Lee-Carter Model equation is given through:

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t} \tag{1}$$

where,

$m_{x,t}$ central rate of death at age $x$ in year $t$.
$a_x$ average of log $[m_x, t]$ across years.
$b_x$ relative speed of change at each age also known as age component.
$k_t$ time-varying parameter.
$\varepsilon_{x,t}$ residual at age $x$ and time $t$.

To get unique solutions, coefficient $a_x, b_x$ and $k_t$ are set $a_x = \frac{1}{T}\sum \ln(m_{x,t})$, $\sum_x b_x = 1$ and $\sum_t b_t = 1$. Then, adopting Singular Value Decomposition (SVD) method to acquire the estimation values of $b_x$ and $k_t$. To estimate the value of $k_t$, Autoregressive Integrated Moving Average Model (ARIMA) is applied. The equation of the ARIMA model is given:

$$k_t = k_{t-1} + \phi_1 k_{t-1} - \phi_1 k_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-1} \tag{2}$$

where $k_t$ is denoted as the actual value at the time period, $t$, $\varepsilon_{x,t}$ denoted as random error at time period, $t$, and $\phi_1$ and $\theta_1$ denoted as parameters of the model.

## 2.3 The CBD Model

The Cairns, Blake and Down (CBD) model proposes a predictor structure with two age-periods. The age modulating parameters were specified as $b_x^{(1)} = 1$ and $b_x^{(2)} = x - \bar{x}$. Where $\bar{x}$ is denoted as the average age. This model assumes no static age function; the population is stationary and no cohort effect. The general equation for the CBD model is given:

$$logit(q_{x,t}) = k_t^{(1)} + k_t^{(2)}(x - \bar{x}) \tag{3}$$

where $q_{x,t}$ is the probability of a person age $x$ will die in $t$ years and $k_t^{(1)}$, $k_t^{(2)}$ denoted as the period effects.

## 2.4 The M8 Model

The M8 model is the third generalization of the CBD model and this model is also based on the adjustment of [13]. Unlike the previous models, this model includes cohort effect in the calculation basis. This model proposed the impact of the effect $\gamma_{t-x}^{(3)}$ for any specific cohort reduces over time instead of remaining constant. The formula of the M8 model is given:

$$logit\left(q_{x,t}\right) = \beta_x^{(1)} k_t^{(1)} + \beta_x^{(2)} k_t^{(2)} + \beta_x^{(3)} \gamma_{t-x}^{(3)} \tag{4}$$

where the equation is derived by taking $\beta_x^{(1)} = 1$, $\beta_x^{(2)} = (x - \bar{x})$ and $\beta_x^{(3)} = (x_c - x)$. For some constant, the $x_c$ in this study is the 80 years old. Therefore, the equation generalizes as:

$$logit\left(q_{x,t}\right) = k_t^{(1)} + \left(x - \bar{x}\right) k_t^{(2)} + (x_c - x)\gamma_{t-x}^{(3)} \tag{5}$$

To overcome the identifiable problems, new constraint is introduced by letting $\sum_{x,t} \gamma_{t-x}^{(3)} = 0$. In fitting the mortality models, namely the Lee Carter, CBD and M8 models, the packages of R programming software, known as *StMoMo* [14] and *Demography* [19] were used.

## 2.5 Evaluations of Mortality Forecasting Models

As defined by the Department of Social Welfare, older persons refer to those who are 60 years and above [19]. Therefore, this study fitted the ages between 60 and 80 years old. However, to prove that the CBD and M8 models are only applicable for higher ages (60–80), the evaluation for all ages (0–80) and lower ages (0–59) were also being carried out from 1966 to 2019 (54 years). Primarily, the data was separated into two parts namely the training and validation, with the percentage of training is 60% (1966–1998) and validation is 40% (1999–2019) from the available data.

**Model Goodness of Fit**

The goodness-of fit test usually used to check the fitted model residuals. Consistent residual patterns specify the model's incompetence for properly defining all of the data features. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two parameters chosen to check the fit-of-data quality. The lower the values of AIC and BIC the better the model fits to the data.

The Akaike Information Criterion (AIC) measures the probability penalty for each additional term that is included in the model. The equation is as follow:

$$AIC = 2k - 2\ln(L) \tag{6}$$

The Bayesian Information Criterion (BIC) is the best criterion to balance between the models' complexity and goodness-of-fit.

$$BIC = 2\ln(Nk) - 2\ln(L) \tag{7}$$

where $k$ denoted as the number of parameters estimated in the model and $L$ is the log-likelihood and $N$ is the number of sample.

**Out-Sample Error Evaluation**
For 40% validation set, Root Mean Square Errors (RMSEs) and Mean Absolute Percentage Errors (MAPEs) are used to test the performance of all three selected models for three different age categories namely all age (0–80), lower age (0–59) and higher age (60–80). The Root Mean Square Errors (RMSEs) is commonly used and able to make an outstanding general purpose error metric for numerical predictions [20]. The general formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{x,i} - y_{x,i})^2} \tag{8}$$

Mean Absolute Errors (MAPEs) are mean or average of the absolute percentage errors of forecasted values. By canceling each other out this approach will prevent the issue of both positive and negative errors. The formula as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(y_{x,i} - \hat{y}_{x,i})^2}{y_{x,i}}\right| \tag{9}$$

where $\hat{y}_{x,i}$ is the prediction central death rates for person age $x$ in year $i$, $y_{x,i}$ is the observed central death rates for person age $x$ in year $i$, and $n$ is the number of sample in year.

## 3 Results and Discussions

This section discusses the results obtained from the selected mortality models used in the study. Due to some missing values from the given data, the interpolation process needs to be carried out in filling up the missing values.

Figure 1 demonstrates the plot for observation data which is Malaysian population pattern of logarithm of death rates according to age and time. Several behaviors are shown respectively for both male and female. As it can be seen, the mortality rates are increasing for both genders in all age's group, however male is slightly thinner compared to female and the presence of volatile accident humps in between the ages of 18–30 years old that is more visible in male than female.

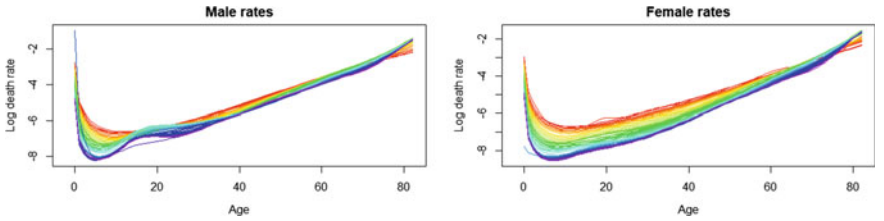



**Fig. 1** Log death rates according to age for Malaysian male and female

## *3.1 Estimation of Models' Parameters*

This section estimates the parameters for all the three models. Figure 2 shows the parameter shapes of the Lee-Carter model. As can be seen, the parameter $a_x$ is increasing by age, whereby the parameter $b_x$ is declining downward as it reacts towards the mortality changes over ages. Meanwhile, the $k_t$, parameter is decreasing by years, with the presence of humps visible for both genders during the year-range (1990–2000). As depicted in Fig. 3, the CBD model shows a declining trend of $kt^{(1)}$ for both genders. However, for mortality-rate dynamic at higher ages $k_t{}^{(2)}$ parameter,

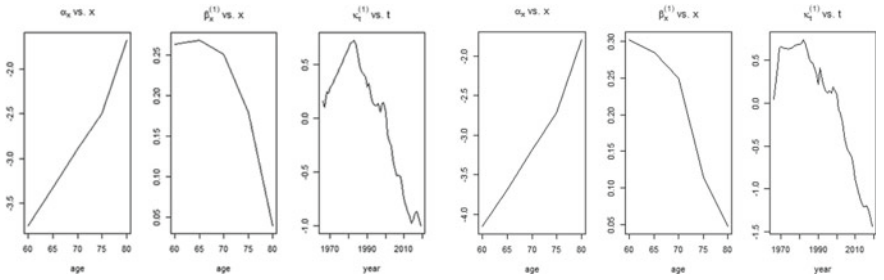



**Fig. 2** Parameters of the Lee-Carter model for Malaysian males (left) and females (right) ages 60–80 over the period of 1966–1998

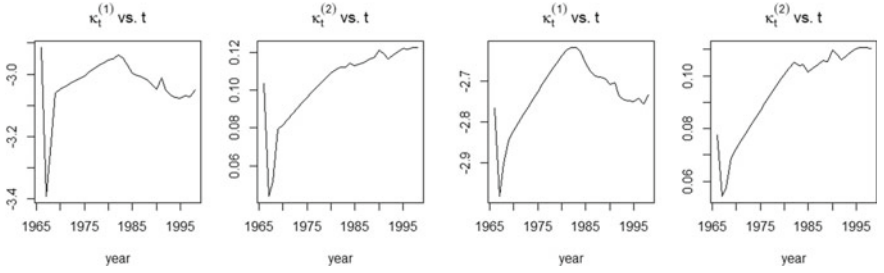



**Fig. 3** Parameters of the CBD model for Malaysian males (left) and females (right) age 60–80 over the period of 1966–1998

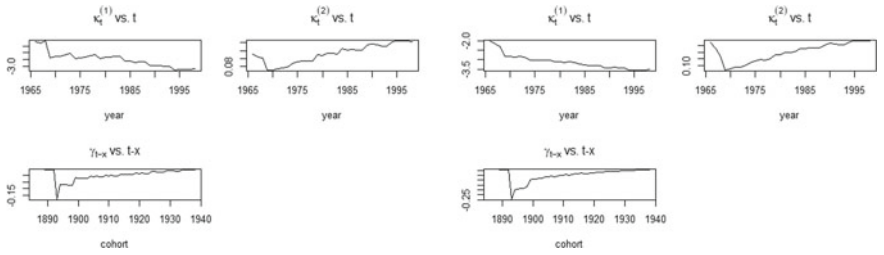



**Fig. 4** Parameters of the M8 model for Malaysian males (right) and females (left) ages 60–80 over the period of 1966–1998

the mortality rates dynamically shoot up at higher ages for both genders. As for the M8 model, Fig. 4 shows that both genders experiencing quite similar trends. More information captured in the M8 model is about its cohort effect, where those who are 80 years old in 1966 data are those who were born in 1886. Hence, the cohort trend captured for both genders are showing improvements in mortality over the years. Nonetheless, both models of the CBD and M8 show the existence of mortality improvement by years.

## Residual Analysis

The residuals were distributed randomly across ages and years. In order to fit the model, this study used scatter diagram to examine the model fitting. The closer the dotted to zero, the better it fits to the line. As can be seen in Fig. 5, the plot of the Lee-Carter model for both genders are scattered and most of the dotted points are far from the fitted lines. For the CBD model, both genders' show dotted lines for certain ages that is extremely too far from the fitted lines. However the patterns for both genders show a quadratic patterns which mean that this model is able to capture the cohort effects and it has been substituted by the additional quadratic effect [9]. Therefore the residuals plot for the M8 as depicted in Fig. 5 has captured the cohort effects as the presence of quadratic patterns are clearly displayed in age residuals for both genders.

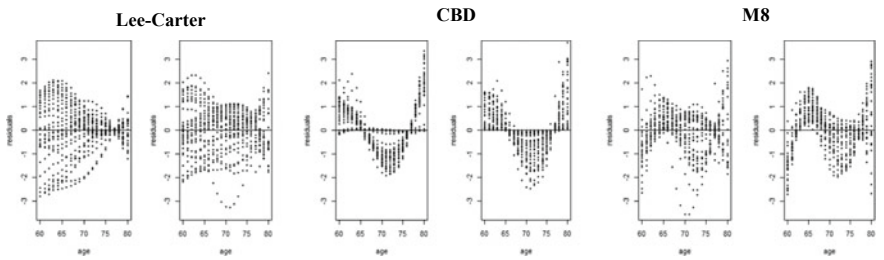



**Fig. 5** The scatter residual diagrams of all the three models for Malaysian males (left) and females (right) age 60–80 over the period of 1966–1998

**Forecasted Parameters for the Selected Models**

Figure 6 shows the patterns of the forecasted graphs of time factors for all three models for higher age category since this study focuses on higher age (60–80). For the Lee-Carter model, it can be seen that, the $k_t^{(1)}$ values for both male and female show the declining pattern across the time. For the CBD model, the parameter $k_t^{(1)}$ shows the declining trend for both male and female, meanwhile for $k_t^{(2)}$ where the mortality is forecasted for higher age, shows the increasing trends for both genders, with narrow values of the forecasted rates. Lastly for the M8 model, the projection trend for period index of $k_t^{(1)}$ and $k_t^{(2)}$ are similar with the CBD model, however the forecasted rates for M8 model is wider as compared to the CBD model. This suggests that the M8 model is more plausible model as compared to the CBD model.

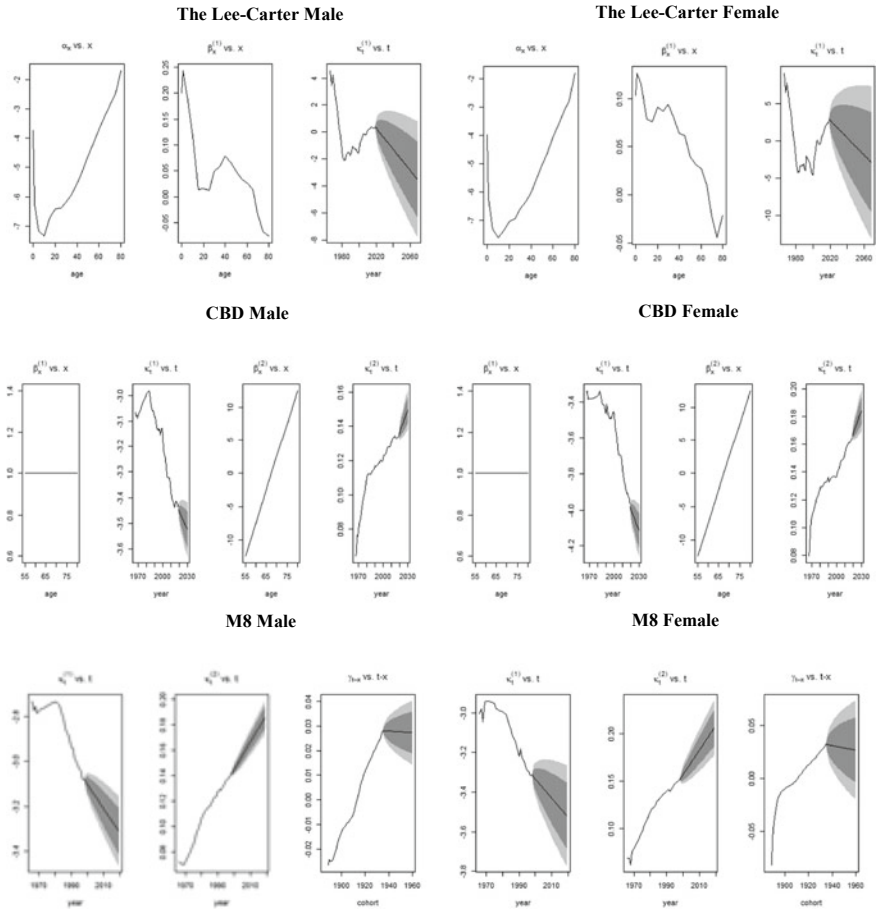



**Fig. 6** The estimated age parameters and estimated and forecasted time components for the all the three models for Malaysia population age 60–80

For the cohort effects, as depicted in the M8 model, the projections for cohort male and female are declining gradually. However, for male, the declining trend is not that significant, the pattern is more towards stagnant.

## *3.2 The Performance of Mortality Models*

**Goodness-of-Fit Analysis**

In this subdivision, the model evaluation for three categories namely all age (0–80), lower age (0–59) and higher age (60–80) were evaluated using AIC and BIC. The lower the AIC and BIC values, the more fit the model to the observation data [21]. Table 1 indicates that, the Lee-Carter model is the best model that fits the data best for all age and lower age categories for both males and females due to the model produced the lowest AIC and BIC. Whereas it shows that the M8 model is the best model to fit the higher age (60–80) data as it gives the lowest values of the AIC and BIC as compared to the other two models.

**Out-Sample Error Analysis**

Table 2 represents the RMSE and MAPE values for all the three models for all age categories. Results show that the Lee-Carter model is the most accurate model to predict mortality rate for all and lower age categories whereas the M8 model outperforms the other two models for high age category. In addition, the result is consistent with [14] that proved the CBD model is more accurate than the Lee-Carter model for older age group.

**Table 1** AIC and BIC for the Lee-Carter, CBD and M8 model according to age categories from year (1966–1998)

| Age range | Model | AIC | | BIC | |
|---|---|---|---|---|---|
| | | Male | Female | Male | Female |
| (0–80) | Lee-Carter | **26,783.32** | **23,750.22** | **27,948.140** | **24,915.04** |
| | CBD | 1,306,643.00 | 1,083,337.00 | 1,307,033.00 | 1,083,728.00 |
| | M8 | 555,230.400 | 375,914.900 | 556,271.10 | 376,955.60 |
| (0–59) | Lee-Carter | **17,090.84** | **14,953.63** | **17,934.15** | **15,796.93** |
| | CBD | 935,653.80 | 725,006.10 | 936,022.40 | 725,374.70 |
| | M8 | 377,893.00 | 237,702.40 | 378,736.30 | 238,545.70 |
| (60–80) | Lee-Carter | 7079.03 | 7409.25 | 6866.68 | 7196.90 |
| | CBD | 7498.19 | 7796.75 | 6802.00 | 7100.56 |
| | M8 | **6258.75** | **6765.39** | **6151.95** | **6658.58** |

**Table 2** RMSE and MAPE for the Lee-Carter, CBD and M8 model according to age categories from year (1999–2019)

| Age range | Model | RMSE | | MAPE | |
|---|---|---|---|---|---|
| | | Male | Female | Male | Female |
| (0–80) | Lee-Carter | **0.050** | **0.061** | **0.038** | **0.036** |
| | CBD | 0.791 | 1.932 | 0.116 | 0.147 |
| | M8 | 6.127 | 2.536 | 0.225 | 0.181 |
| (0–59) | Lee-Carter | **0.072** | **0.085** | **0.032** | **0.032** |
| | CBD | 0.535 | 1.318 | 0.077 | 0.099 |
| | M8 | 14.502 | 14.889 | 0.578 | 0.518 |
| (60–80) | Lee-Carter | 0.242 | 0.0735 | 0.265 | 0.078 |
| | CBD | 0.234 | 0.0699 | 0.254 | 0.073 |
| | M8 | **0.194** | **0.0529** | **0.217** | **0.062** |

## 4 Conclusions

This study compares the performance of the Lee-Carter model and its extension models namely the CBD model and M8 model for Malaysia population from the year 1966–2019. This study is focusing on higher age range of 60–80 years old population in Malaysia. In general, the performance of the M8 model outshined the other two models in every aspect of analyses particularly for higher age category. The performance of the M8 model is not only able to capture the cohort effects, but through a series of analyses in fitting the models, this model proves to fit most analyses best. As for the validation analysis, the M8 model also outperformed the other two models as this model has potential to do the forecasting more precisely. This research proved that the M8 model fit Malaysian data best as compared to other mortality models. The Malaysia data is suitable for the M8 model that include cohort effect to forecast the mortality rate especially for higher ages category. Nevertheless, based on the out-sample error for all the separate age categories namely, overall, lower and higher ages, the M8 model only significant for higher ages category but not for overall and the lower ages category.

**Acknowledgements** The author would like to acknowledge the supports from the Ministry of Education (MOE) and Research Management Centre, Universiti Teknologi MARA (UiTM) for the financial support through Fundamental Research Grant Scheme with file number of 600-IRMI/FRGS 5/3 (125/2019) and for the permission to publish this research to a journal or conference proceeding.

## References


1. Ngataman N, Ibrahim RI, Yusuf MM (2016) Forecasting the mortality rates of Malaysian population using Lee-Carter method. In: AIP conference proceedings, vol 1750, no 1. AIP Publishing LLC, Melville, p 020009
2. Department of Statistics Malaysia press release abridged life tables, Malaysia, 2017–2019 (2019), p 4. Retrieved 23 Dec 2019 from https://www.dosm.gov.my/
3. Chan MF, Kamala Devi M (2015) Factors affecting life expectancy: evidence from 1980–2009 data in Singapore, Malaysia, and Thailand. Asia Pac J Public Health 27(2):136–146
4. McGarry KM (2020) Perceptions of mortality: individual assessments of longevity risk. Wharton Pension Research Council working paper
5. Kamaruddin HS, Ismail N (2018) Forecasting selected specific age mortality rate of Malaysia by using Lee-Carter model. Paper presented at the journal of physics: conference series
6. Lee RD, Carter LR (1992) Modeling and forecasting US mortality. J Am Stat Assoc 87(419):659–671
7. Shair SN, Zolkifi NA, Zulkefi NF, Murad A (2019) A functional data approach to the estimation of mortality and life expectancy at birth in developing countries. Pertanika J Sci Technol 27(2) (2019)
8. Pascariu MD, Canudas-Romo V, Vaupel JW (2018) The double-gap life expectancy forecasting model. Insur Math Econ 78:339–350
9. Seklecka M, Md. Lazam N, Pantelous AA, O'Hare C (2019) Mortality effects of economic fluctuations in selected Eurozone countries. J Forecast 38(1):39–62
10. Zili AHA, Mardiyati S, Lestari D (2018) Forecasting Indonesian mortality rates using the Lee-Carter model and ARIMA method. In: AIP conference proceedings, 2023, no 1, p 020212
11. Booth H, Maindonald J, Smith L (2002) Applying Lee-Carter under conditions of variable mortality decline. Popul Stud 56(3):325–336
12. Cairns AJ, Blake D, Dowd K (2006) A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. J Risk Insur 73(4):687–718
13. Renshaw AE, Haberman S (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. Insur Math Econ 38(3):556–570
14. Cairns AJ, Blake D, Dowd K, Coughlan GD, Epstein D, Ong A, Balevich I (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North Am Actuarial J 13(1):1–35
15. Millossovich P, Villegas AM, Kaishev VK (2018) Stmomo: an r package for stochastic mortality modelling. J Stat Softw 84(3)
16. Rabi A, Mansor N, Awang H, Kamarulzaman ND (2019) Longevity risk and social old-age protection in Malaysia: situation analysis and options for reform
17. Zulkifle H, Yusof F, Nor SRM (2019) Comparison of Lee Carter model and Cairns, Blake and Dowd model in forecasting Malaysian higher age mortality. MATEMATIKA 35(4):65–77
18. United Nations (2019) World population prospects: the 2019 revision. Department of Economics and Social Affairs, Population Division. Retrieved 3 Mar 2020 from https://population.un.org/wpp/Download/Standard/Population/
19. Hyndman MRJ (2012) Package 'demography'
20. Lazim MA (2012) Introductory business forecasting: a practical approach, 2nd edn. Penerbitan UiTM, Shah Alam, Selangor
21. Maccheroni C, Nocito S (2017) Backtesting the Lee-Carter and the Cairns–Blake–Dowd stochastic mortality models on Italian death rates. Risks 5(3):34

# Assessing Python Programming Through Personalised Learning Styles Model


**Sin-Ban Ho, Sek-Kit Teh, Ian Chai, Chuie-Hong Tan, Swee-Ling Chean, and Nur Azyyati Ahmad**

**Abstract** Learning styles, cognitive traits, personality, and learning preferences can vary greatly. That is why there is a great variety in how people receive and process information. Personalizing learning materials according to learner's learning styles could enhance learner's learning motivation and lead to better learning performance. This paper examines the relationship between learner's learning styles and learning performance by proposing three different sets of documentation to test the relationship between the two learning styles of Felder-Silverman and learning performance. To test the proposed documentations and hypotheses, 182 participants in Multimedia University, Cyberjaya, Malaysia answered the Index of Learning Styles (ILS) questionnaire by Felder-Silverman and participated in a documentation experiment in Python programming. The data gathered was analysed using statistical Chi-square test. The results showed that learning performance was enhanced when the documentation was provided in a learning style that matched the subject's learning style. The confirmed personalised learning styles model can be beneficial to teachers and

S.-B. Ho (✉) · S.-K. Teh · I. Chai · N. A. Ahmad
Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
e-mail: sbho@mmu.edu.my

S.-K. Teh
e-mail: nicholas.teh93@gmail.com

I. Chai
e-mail: ianchai@mmu.edu.my

N. A. Ahmad
e-mail: azyyati.ahmad@mmu.edu.my

C.-H. Tan
Faculty of Management, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
e-mail: chtan@mmu.edu.my

S.-L. Chean
Lee Kong Chian Faculty Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia
e-mail: karenchean@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_13

e-learning recommendation systems when they provide students with materials that are personalised.

**Keywords** Knowledge management · Knowledge discovery · Web-based computing · Personalisation

## 1 Introduction

Many methods have been studied and applied in the presentation of knowledge or information to beginners. Learning had previously taken place in a given place and time. Teachers had often been seen as the primary source of new information [1, 2]. Nonetheless, this strategy is facing difficulties because, for example, the world is evolving, increasing the number of applicants, new enrolment from various countries, expanded penetration of research areas and Internet development adds to the challenge. In comparison to these problems, the academic experience varies dramatically from generation to generation.

Students now have faster access and multiple ways to search for information on the Internet with the invention of the Internet and the creation of the World Wide Web (WWW). Therefore, users can exchange information more easily than ever before. To newcomers, improved dissemination of learning around the world means that people can better themselves with schooling as it can boost their living standard and socio-economic position. Technology development and information access raise the number of knowledgeable people, posing a problem for government organizations. They need to ensure that there is enough space to gather information and educate these people to achieve the goal of lifelong learning. Other challenges, such as geographic separation, lack of accessible places, and time constraints, require researchers to consider other knowledge delivery approaches. Current knowledge-based solutions could not address the complexities and limitations.

Different learning strategies were designed to overcome the challenges and limitations faced by learners, instructors and universities. A new approach to learning has been established with a steady increase in Internet speeds, the minimalism introduced with Web 2.0 standards and wider accessibility. This modern learning methodology is also called e-learning. This technique for the distribution of information reduces geographic isolation, time constraints and restrictions on location.

## 2 Background Study

Different learners have unique characteristics, for instance learning styles, learning preferences and personalities. Every individual acquires and processes information in a different way. Personalizing learning materials according to learner's learning styles could enhance learner's learning motivation and contribute to a better performance.

The process of learning and acquiring knowledge is a complicated and challenging process. A few factors such as acquiring and processing of knowledge by learners in terms of their common knowledge, developmental characteristics, and environmental components has an important part to play in this process. The learning process is influenced by various factors that will present different challenges to learners. There are several results from different research studies show that taking these differences into account while creating learning and teaching settings contributes to the increase of the effectiveness of learning activities, and efficiency in class [3–5].

The learning needs of the students can be addressed when considering their learning interests and demands [5]. The integration of information and communication technology (ICT) into educational settings has also contributed significantly to learning methods [6, 7]. This technology has driven developments in e-learning settings and their personalization according to learner's knowledge-acquisition needs.

An individual's preferred method of learning can be determined by first identifying the individual's learning style as learning styles describe learner's attitudes and actions when it comes to learning. Learning styles are crucial in educational environments as it may support students and teachers to become more self-conscious of their own strengths and weaknesses [8]. Learning styles are also one of the most vital factors to be utilised for taking into account individual differences [9].

Learning styles are the learning patterns and variations of an individual [10–12]. Numerous research studies investigate the efficiency and productivity of learning settings based on different individual learning styles. Such research studies indicate that the learning process in environments appropriate for learning styles has a positive impact on students' memory and application of information to a particular course or subject [6]. In addition, other empirical studies have shown that learning environments based on learning styles have a significant impact on the results or success of students performance [13–15].

## 3 Methodology

The research objectives of this paper are:

- To investigate the impact of student's learning styles and their performance in an introductory programming course.
- To propose a method in using information of student's learning styles as a guide in personalizing student's learning materials delivery approaches and study habits in the learning of programming.
- To evaluate the proposed method for the design of learning strategies.

This study applied an exercise-based experiment. These experiments were conducted with undergraduates from the Faculty of Computing and Informatics as participants. The authors did not inform the experimental participants about

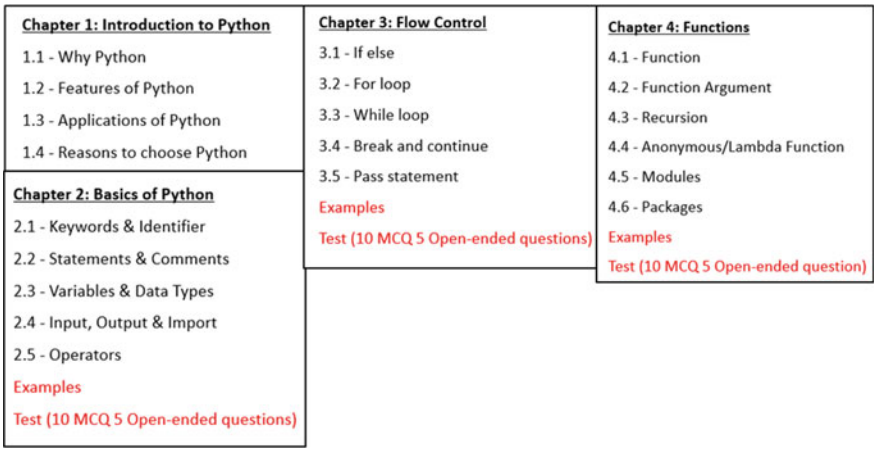



**Fig. 1** Python learning syllabus

the research goals. The materials, examples and test questions are adapted from Schneider [16], as shown in Fig. 1.

An e-learning system has been designed and coded using Microsoft VB.NET and ASP.NET to automate the process of assessing students' pre-dominant learning styles using the Felder-Silverman model [12] and the personalization of students' learning materials. After completion, the system was configured and deployed in Microsoft Azure (cloud service) as shown in Fig. 2.

This system is coded to automate the process of assessing students' pre-dominant learning styles using the Felder-Silverman model, personalise of students' learning materials and record students' learning performance. A different experimental setting is developed for this part, which was chosen in order to compare student's learning styles and learning preferences. After the learning styles are assessed, the system will personalise the learning materials according to four different documentation styles, namely Verbal/Sequential, Verbal/Global, Visual/Sequential, and Visual/Global. Figure 3 shows the two sets of learning styles (Visual/Verbal and Sequential/Global) tested in this series of experiments. The participants can attempt to do the examples at the end of each chapter or sub-chapter. After that, they need to attempt a test at the end of each chapter or sub-chapter.

This experiment involves 182 Computer Science undergraduates at Multimedia University (MMU) Cyberjaya, which were categorised into three different documentation groups. The groups comprise of personalised learning style (*pLS*) group, opposite learning style (*oppLS*) group and control group (*ctrlGrp*). For the *pLS* group, learning works best when students are instructed in their preferred learning style. We would like to investigate whether the best for people with one learning style might not work so well with people in the *oppLS* group with different learning style. Finally, the control group (*ctrlGrp*) underwent the traditional method of learning materials given to them without involving any digital usage of e-learning system. Upon using

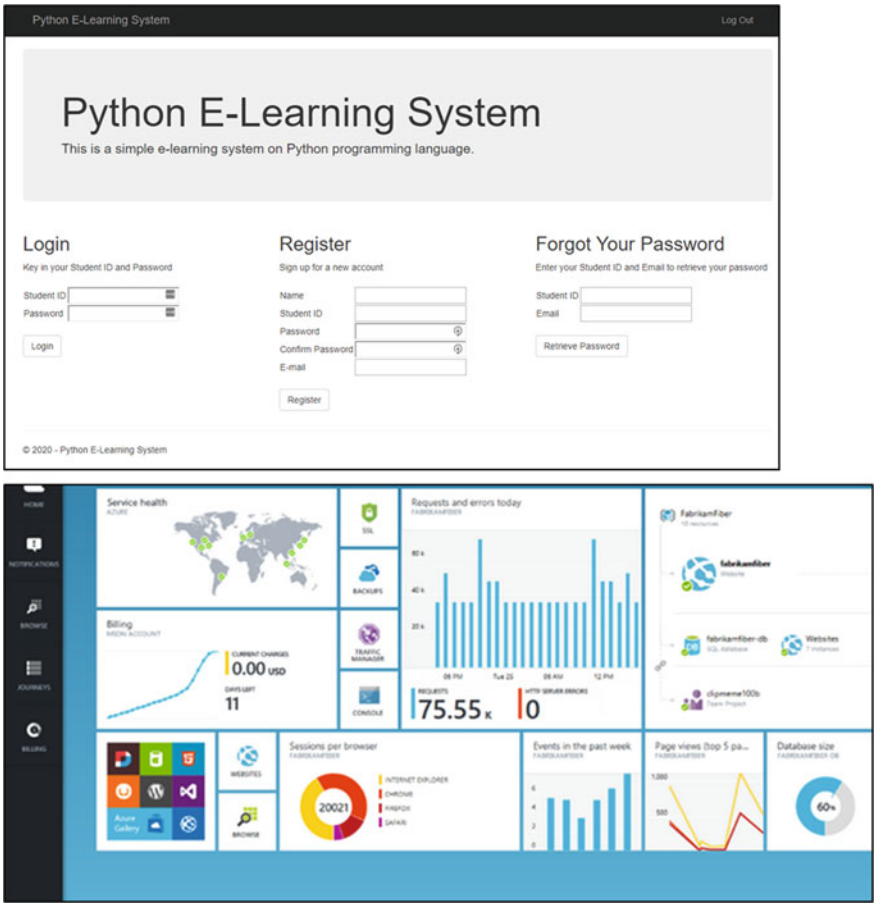



**Fig. 2** Snapshots of e-learning system deployed in Microsoft Azure

the manual materials, the control group would attempt the exercise without considering any of their learning styles at all. In summary, our hypothesis is summarised as follows: *H0*—There is no difference among all three documentation groups (*pLS*, *oppLS*, *ctrlGrp*) for the participants in performing the given Python exercise.

## 4 Results and Discussions

Analysis of data obtained with the student's learning styles to identify some possible patterns and verify if there is some correlation between the participants' learning styles and their performance. In addition, the analysis of data could also help to evaluate whether the method used in this research is feasible in the design of student's

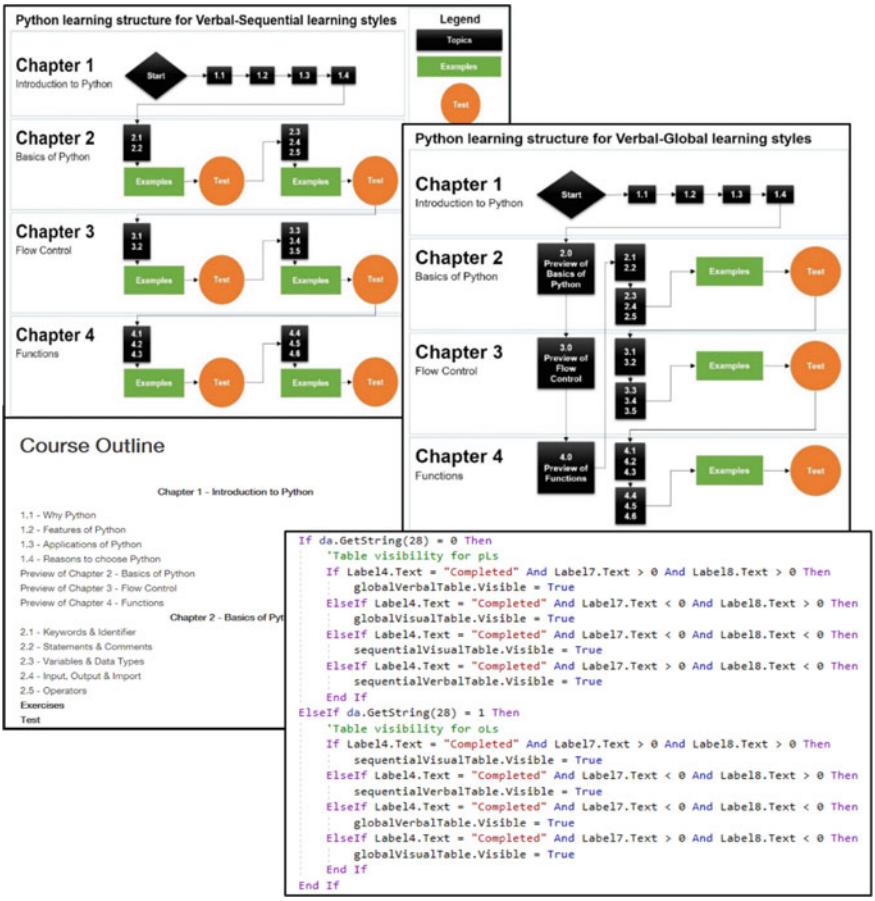



**Fig. 3** Snapshots of the e-learning system

learning strategies. To assess student's performance, this research uses indicators of completion time (time taken to complete a test), and comprehension (understanding of a code). We performed a statistical analysis of the 182 responses obtained through the Statistical Package for Social Science (SPSS). The dependent variables of all three Stages (Stage 1 [Chapter 2], Stage 2 [Chapter 3], and Stage 3 [Chapter 4]) are:

(a) Time taken to complete a test (*complTimeStg1*, *complTimeStg2*, *complTimeStg3*).
(b) Comprehension in answering multiple choice questions (*comprMcqStg1*, *comprMcqStg2*, *comprMcqStg3*)
(c) Comprehension in answering structured questions (*comprStrucStg1*, *comprStrucStg2*, *comprStrucStg3*).

Table 1 shows the test results for normality for these dependent variables. Three

**Table 1** Results of Kolmogorov–Smirnov normality test

| Category | *p*-value | Category | *p*-value | Category | *p*-value |
|---|---|---|---|---|---|
| 1. *complTimeStg1* | 0.123 | 4. *complTimeStg2* | 0.219 | 7. *complTimeStg3* | 0.194 |
| 2. *comprMcqStg1* | 0.002* | 5. *comprMcqStg2* | 0.014* | 8. *comprMcqStg3* | 0.000* |
| 3. *compStrucStg1* | 0.008* | 6. *compStrucStg2* | 0.000* | 9. *compStrucStg3* | 0.000* |

*Statistically significant at 0.050 level (with $p < 0.050$)

dependent variables are normally distributed, (*complTimeStg1*, *complTimeStg2*, *complTimeStg3*), with *p*-values more than 0.050. Therefore, the median is used for other dependent variables instead of the mean.

Table 2 shows the bold-faced cells having dependent variables with higher (mean/median) scores. Three dependent variables (*complTimeStg1*, *complTimeStg2*, *complTimeStg3*) are normally distributed, hence the median is used for the other dependent variables instead of the mean. Each documentation group (*pLS*, *oppLS* and *ctrlGrp*) has different numbers of participants because each group was assigned according to participants' lab classes by the university.

**Table 2** The categories descriptive statistics

| | **Mean** | | | **Std. dev** | | |
|---|---|---|---|---|---|---|
| **Dependent variable (sample size, *n*)** | ***pLS* (76)** | ***oppLS* (40)** | ***ctrlGrp* (66)** | ***pLS*** | ***oppLS*** | ***ctrGrp*** |
| 1. *complTimeStg1* (hh:mm:ss) | **0:14:37** | 0:17:37 | 0:27:51 | 0:09:21 | 0:14:42 | 0:12:15 |
| 2. *complTimeStg2* (hh:mm:ss) | **0:32:00** | 0:48:41 | 1:01:34 | 0:18:20 | 0:24:25 | 0:10:00 |
| 3. *complTimeStg3* (hh:mm:ss) | **0:20:55** | 0:25:08 | 0:57:19 | 0:11:27 | 0:10:11 | 0:10:11 |
| | **Median** | | | **Std. dev** | | |
| | ***pLS*** | ***oppLS*** | ***ctrlGrp*** | ***pLS*** | ***oppLS*** | ***ctrlGrp*** |
| 4. *comprMcqStg1* (scale: 0–10) | **8.00** | 6.00 | 7.00 | 1.363 | 1.207 | 1.233 |
| 5. *comprStrucStg1* (scale: 0–10) | **9.00** | 7.50 | 7.00 | 1.519 | 2.444 | 1.484 |
| 6. *comprMcqStg2* (scale: 0–10) | 7.00 | 7.00 | 6.00 | 1.467 | 1.476 | 1.388 |
| 7. *comprStrucStg2* (scale: 0–10) | **10.00** | 8.50 | 8.00 | 1.285 | 2.262 | 1.784 |
| 8. *comprMcqStg3* (scale: 0–10) | 8.00 | 8.00 | 8.00 | 1.648 | 1.318 | 1.115 |
| 9. *comprStrucStg3* (scale: 0–10) | 9.00 | 9.00 | **10.00** | 1.736 | 1.748 | 1.686 |

### 4.1 Completion Time for Stage 1, Stage 2 and Stage 3

Some items are bold-faced in Table 2 to show that a particular group performs better than the other two groups. For example, the personalised learning group in *complTimeStg1* took 14 min 37 s to complete the exercise in terms of completion time. The opposite learning group, meanwhile, took a longer period of 17 min 37 s and it took 29 min 51 s for the control learning group to perform the same exercise. Furthermore, in *complTimeStg2*, the personalised group completed the fastest. Students in the personalised learning group only took 32 min to complete the given exercise whereas the opposite learning group completed in 48 min 41 s and control learning group finished in 1 h 1 min 34 s. Finally, in terms of completion time for Stage 3, *complTimeStg3*, students in the personalised learning group completed faster in 20 min 55 s as compared to the opposite learning group, 25 min 8 s and the control group, 57 min 19 s.

### 4.2 Comprehension

As for comprehension in answering multiple choice questions and structured questions in the given exercise to students in all three stages, *comprMcqStg1*, the personalised learning group has the highest median. Next, let us consider the comprehension in answering multiple choice questions and structured questions in the specified exercise to students in all three stages. For *comprMcqStg1*, the personalised learning group has the highest median of 8.00 correct answers (out of 10), as compared to the opposite learning group, which has a median of 6.00 correct answers, and the control learning group, which has a median of 7 correct answers. In comparison, for *comprStrucStg1*, the personalised learning group has a median of 9.00 correct answers, while the opposite learning group has a median of 7.50 correct answers, and the control learning group has a median of 7.00 correct answers. The rubric used for the scale 0–10 were based on the exercises extracted from the Python practices evaluation scheme [16]. For each variable assessed, ten coding questions are formulated. Each correct Python answer contributes to one unit of score into the scale of 0–10.

For Stage 2, *comprMcqStg2*, both the personalised learning group and the control learning group have an average of 7.00 correct answers, while the opposite learning group has an average of 6.00 correct answers. Regarding *comprStrucStg2*, the personalised learning group has the highest median value of 10.00 correct answers as compared to the opposite learning group, which has a median of 8.50 and the control learning group, which has a median of 8.00 correct answers.

All three learning groups, personalised learning group, opposite learning group and control learning group has the same median value of 8.00 correct answers for *comprMcqStg3*. Lastly, as for *comprStrucStg3*, the control learning group has the

**Table 3** Multivariate effects on dependent variables

| Category | *F* | *p*-value |
|---|---|---|
| 1. *complTimeStg1* | 12.730 | **0.000*** |
| 2. *complTimeStg2* | 25.512 | **0.000*** |
| 3. *complTimeStg3* | 140.222 | **0.000*** |

*Statistically significant at 0.050 level with $p < 0.050$ (2-tailed)

highest median of 10.00 correct answers whereas the personalised learning group and opposite learning group only has a median of 9.00 correct answers.

## *4.3 Significance Among the Three Documentation Groups*

Table 3 shows the results of the separate multivariate tests. These F-tests are performed to indicate the specific dependent variables that are important across the three different learning groups. The *p*-values are derived by MANOVA (Multi-variate Analysis of Variance) testing of results between subjects. These results imply high significance differences in mean scores through Wilks' Lambda = 0.651, $F(6,394) = 15.708$ ($p < 0.0001$).

With respect to *complTimeStg1*, *complTimeStg2*, and *complTimeStg3* in Table 2, participants from the personalised learning (*pLS*) group complete their entire task faster than the opposite learning (*oppLS*) group and the control learning group. When the standard significance level of 0.050 (95 percent probability) is found in Table 3, the personalised learning group provides evidence that *complTimeStg1*, *complTimeStg2*, and *complTimeStg3* are much quicker. The personalised learning group participants are significantly faster than the opposite and the control learning groups.

The non-parametric Mann–Whitney test is used because the six dependent variables (*comprMcqStg1, comprStrucStg1, comprMcqStg2, comprStrucStg2, comprMcqStg3*, *comprStrucStg3*) are not normally distributed over the comparison of the three learning groups. Table 4 shows that *complTimeStg1*, *complTimeStg2*,

**Table 4** Mann–Whitney test results on the learning groups

| | Mean rank | | | | |
|---|---|---|---|---|---|
| Categories | *pLS* | *oppLS* | *ctrlGrp* | $\chi^2$ | *p*-value |
| 1. *comprMcqStg1* | 76.16 | 41.95 | 63.39 | 20.852 | **0.000*** |
| 2. *comprStrucStg1* | 76.34 | 49.11 | 56.05 | 13.629 | **0.001*** |
| 3. *comprMcqStg2* | 68.43 | 64.66 | 48.41 | 7.799 | **0.020*** |
| 4. *comprStrucStg2* | 77.59 | 56.08 | 47.84 | 16.718 | **0.000*** |
| 5. *comprMcqStg3* | 70.23 | 55.10 | 54.54 | 5.631 | 0.060 |
| 6. *comprStrucStg3* | 54.24 | 58.20 | 69.06 | 4.330 | 0.115 |

*Statistically significant at 0.050 level with $p < 0.050$ (2-tailed)

*complTimeStg3*, *comprStrucStg2* with $p$-values < 0.050 have significant differences among the three learning groups. However, *comprMcqStg3* and *comprStrucStg3* with $p$-values greater than 0.05, have no significant difference among the three learning groups. For the advanced Stage 3, the participants performed well to complete the given task, irrespective of which type of documentation was given to them.

In Table 4, with respect to *comprMcqStg1*, *comprStrucStg1*, *comprMcqStg2* and *comprStrucStg2*, participants from the personalised learning group show significantly better results than those from the opposite and the control learning groups in the early stages. This therefore follows the rejection of the H0 hypothesis in Sect. 3 for these variables. Such rejection means that in facilitating learning to the learners, the personalised and control learning group was distinct. As noticed by Ho and Tan [17], most undergraduates also come from the sequential learning style. As such, the sequential documentation style suits most intermediate students, who usually have a sequential learning style. These results support the personalization of learning styles that can be beneficial to teachers and e-learning in consistent with the previously published works [18–23]. The personalized materials according to students' learning styles establish significant improved comprehension in both the multiple choices and structured responses as shown in Table 4.

## 5 Conclusion

In conclusion, this paper provides the following three major contributions.

- A Python introduction technique was proposed to cater to four different learning style groups namely Verbal/Sequential, Verbal/Global, Visual/Sequential and Visual/Global. Results from the two series of experiments conducted in this research demonstrated that students participating in personalised learning environments are more motivated and tend to complete faster than those in a traditional learning environment. Lecturers can benefit from this Python introduction technique especially in educating students in an introductory programming course.
- Next, an assessment methodology was designed for the recognition of learning styles that will help lecturers to identify suitable methods in teaching. This assessment methodology presents frameworks for lecturers or teachers to prepare students' learning materials for different learning groups, for example, the personalised learning group (*pLS*), the opposite learning group (*oppLS*) and the control learning group (*ctrlGrp*).
- The results from this series of experiment provide ways or options for lecturers or teachers to develop their learning strategies. Knowing the learning styles of each learner can help lecturers or teachers to identify students' learning preferences and strengths, which can be utilised in instructional designs as to improve the students' learning performance.

To date, limited research has been conducted to improve learning experience and academic achievement by integrating students' learning styles in their learning

process. It provides some key ideas to the existing literature in improving performance of learning programming. The results of this paper have also contributed to the knowledge and literature in educational research. To reiterate, this research aims to use the assessment of learning styles to improve learning in programming by developing a method for learning programming, particularly in Python. The following presents conclusions on the findings to the research objectives and research questions.

Firstly, different people have different learning styles, which can change the way they learn [18] and performance in different situations. For this reason, how we present new materials to students can change how well people learn. We have shown that, in the field of learning programming, the students' learning style can influence how they perceive the materials given to them. Therefore, it motivates us to figure out the students' learning styles, and present the information to them in a manner that is more suitable to them, so that they can learn more efficiently. These findings examined how multiple learning styles affect how well people learn programming in order to propose ways for teachers to develop their materials.

Secondly, e-learning services are not new anymore, but tailoring them in such a way as to help different types of learners is still a challenge. For this reason, we have proposed a way to personalise learning materials on an e-learning system according to students' learning styles. This is well supported from the concept of personalised learning styles model literature [19–23]. The purpose of the e-learning is to help learners to accomplish their learning objectives. Because learning styles theory suggests that how difficult it is for someone to learn something new could be greatly influenced by whether the materials they are presented with matches their learning style or not, the idea of personalisation is very attractive. This is why it makes sense to assess students and trainees' learning styles and choosing to present materials in a way that matches that. For this reason, it has become more common for lesson plans include a plan for how to address students with different learning styles. That is why our findings are relevant to educational theory and practice.

Thirdly, the proposed method for the design of learning strategies is assessed. Pertaining the results shown in the series of experiments in this paper, the participants from the personalised learning group (*pLS*) completed their entire task faster than the other learning groups, namely the opposite learning group (*oppLS*) and the control learning group (*ctrlGrp*). It is hypothesised that providing instruction based on individuals' preferred learning styles improves learning and this can be incorporated into the design of learning strategies by lecturers or teachers.

In the future, we may try a course on advanced topics and analyse it with Structural Equation Modelling (SEM). SEM is a multivariate statistical technique, which can simultaneously analyse a series of dependent relationships [24]. SEM allows the evaluations of a single model containing all relationships in a hypothesis. This could be more accurate than trying to analyse each way of learning programming individually.

**Acknowledgements** The authors would like to thank the financial support given by the Fundamental Research Grant Scheme, FRGS/1/2019/SS06/MMU/02/4 and Multimedia University, Cyberjaya, Malaysia (Project ID: MMUE/190031).

## References


1. Hoyos AAC, Velásquez JD (2020) Teaching analytics: current challenges and future development. IEEE J Lat Am Learn Technol (IEEE-RITA) [IEEE Revista Iberoamericana de Tecnologias del Aprendizaje] 15(1):1–9
2. Lye SY, Koh JHL (2014) Review on teaching and learning of computational thinking through programming: what is next for K-12? Comput Hum Behav 41:51–61. https://doi.org/10.1016/j.chb.2014.09.012
3. Ho SB, Chean SL, Chai I, Tan CH (2019) Engineering meaningful computing education: programming learning experience model. In: Proceedings of 2019 IEEE international conference on industrial engineering and engineering management (IEEM), 15–18 Dec 2019, Macao. IEEE, pp 925–929
4. Shanmugam L, Nadesan G (2019) An innovative module for learning computational thinking skills among undergraduate students. Int J Acad Res Progressive Educ Dev 8(4):116–129
5. Ogange BO, Agak JO, Okelo KO, Kiprotich P (2018) Student perceptions of the effectiveness of formative assessment in an online learning environment. Open Praxis 10(1):29–39
6. Clark RM, Dickerson SJ (2018) A case study of post-workshop use of simple active learning in introductory computing sequence. IEEE Trans Educ 61(3):167–176
7. Chean SL, Ho SB, Chai I (2018) A conceptual framework on constructing effective learning content for programming novices. In: Proceedings of international conference on informatics, computing & applied mathematics (ICICAM2017), 7–9 Oct 2017, UniSEA (Universiti Sultan Zainal Abidin), Kuala Nerus, Terengganu, Malaysia. Int J Eng Technol (UAE) 7(2.15):150–153
8. Lacave C, Molina AI, Redondo MA (2018) A preliminary instrument for measuring students' subjective perceptions of difficulties in learning recursion. IEEE Trans Educ 61(2):119–126
9. Ho SB, Teh SK, Chan GY, Chai I, Tan CH (2018) Sequential and global learning styles as pathways to improve learning in programming. In: Proceedings of the 4th international conference on computational science & technology (ICCST2017), 29–30 Nov 2017, ParkRoyal Hotel, Kuala Lumpur, Malaysia. Lecture notes in electrical engineering (LNEE) 488, pp 1–10
10. Ho SB, Chai I, Tan CH (2016) Different styles for different complexity: empirical findings of documentation styles for information technology management. In: International conference on information in business & technology management (I2BM'2016), 26–28 Jan 2016, Melaka, Malaysia. Inf J 19(7(A)):2643–2648
11. Ho SB, Chai I, Tan CH (2014) A comparison of three documentation styles for educational data analysis. In: Herawan T, Deris M, Abawajy J (eds) Proceedings of the first international conference on advanced data and information engineering (DaEng-2013). Lecture notes in electrical engineering 285. Springer, Singapore, pp 703–710
12. Felder RM, Spurlin J (2005) Applications, reliability and validity of the index of learning styles. Int J Eng Educ 21(1):103–112
13. Lockwood J, Mooney A (2018) Computational thinking in secondary education: where does it fit? A systematic literary review. Int J Comput Sci Educ Schools 2(1):41–60
14. Ibriwesh I, Ho SB, Chai I, Tan CH (2019) Prioritizing solution-oriented software requirements using the multiple perspective prioritization technique algorithm: an empirical investigation. Concurrent Eng Res Appl (CERA) 27(1):68–79. https://doi.org/10.1177/1063293X18808559
15. Ibriwesh I, Ho SB, Chai I (2018) Overcoming scalability issues in analytic hierarchy process with ReDCCahp: an empirical investigation. Arab J Sci Eng (AJSE) 43(12):7995–8011. https://doi.org/10.1007/s13369-018-3283-2
16. Schneider DI (2016) An introduction to programming using python, Global edn. Pearson Education Limited, Essex, England
17. Ho SB, Tan CH (2015) Local population: a study in the influence of learning styles in computing field. In: International conference on local government (ICLG-2015), 5–6 June 2015, Langkawi, Kedah, Malaysia. Aust J Basic Appl Sci 9(22):1–7
18. Zacharis NZ (2011) The effect of learning style on preference for web-based courses and learning outcomes. Br J Educ Technol 42(5):790–800

19. Nabizadeh AH, Leal JP, Rafsanjani HN, Shah RR (2020) Learning path personalization and recommendation methods: a survey of the state-of-the-art. Expert Syst Appl 159(113596):1–20
20. Sanjabi T, Montazer GA (2020) Personalization of e-learning environment using the Kolb's learning style model. In: 2020 6th international conference on web research (ICWR), 22–23 Apr 2020, Tehran, Iran. IEEE, pp 89–92
21. Yi B, Zhang D, Wang Y, Liu H, Zhang Z, Shu J, Lv Y (2017) Research on personalized learning model under informatization environment. In: 2017 international symposium on educational technology (ISET), 27–29 June 2017, Hong Kong. IEEE, pp 48–52
22. Hasibuan MS, Nugroho L, Santosa P (2018) Prediction learning style based on prior knowledge for personalized learning. In: 2018 4th international conference on science and technology (ICST), 7–8 Aug 2018, Yogyakarta, Indonesia. IEEE, pp 1–5
23. Lei G, Luo X, Yang S, Xiao K (2021) Adaptive online learning model based on big data. In: Sugumaran V, Xu Z, Zhou H (eds) Application of intelligent systems in multi-modal information analytics (MMIA 2020). Advances in intelligent systems and computing, 1233. Springer, Cham, pp 643–649. https://doi.org/10.1007/978-3-030-51431-0_92
24. Hair JF, Black WC, Babin BJ, Anderson RE (2019) Multivariate data analysis, 8th edn. Cengage Learning EMEA, Andover, Hampshire, UK

# The Programming Learning Assessment Model for Measuring Student Performance


**Swee-Ling Chean, Sin-Ban Ho, Ian Chai, Chuie-Hong Tan, Sek-Kit Teh, and Nur Azyyati Ahmad**

**Abstract** With recent pandemic, many students cannot join the class in physical classroom. The needs for e-learning and self-assessment become more salient than before. The teaching mode has been changing from teacher-centered to student-centered method. E-learning environment is practically a highly essential software application in the education field. However, programming-specific functionalities are hardly to be found on most of the general-purpose learning platforms, which may be unwieldy and unnecessarily complex to instructors and students in the programming learning process. This research aims to design a self-assessment model for a better support of programming e-learning, especially with exist of mandatory programming-specific functionalities. It's believed that student background and effort have close correlation with their programming performance. More data to verify the correlations associated with positive learning outcome. In this research, we highlight the relationship between student background and student performance levels for introducing personalised self-assessment sets for students to learn programming. We propose and discuss **L**anguage, **E**ducation, **A**chievement,

S.-L. Chean
Lee Kong Chian Faculty Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia
e-mail: karenchean@gmail.com

S.-B. Ho (✉) · I. Chai · S.-K. Teh · N. A. Ahmad
Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
e-mail: sbho@mmu.edu.my

I. Chai
e-mail: ianchai@mmu.edu.my

S.-K. Teh
e-mail: nicholas.teh93@gmail.com

N. A. Ahmad
e-mail: azyyati.ahmad@mmu.edu.my

C.-H. Tan
Faculty of Management, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
e-mail: chtan@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_14

and **P**rogramming (LEAP) and **P**rogramming **L**earning **A**ssessment (PLA) models to fill in the gap between the background knowledge and student competencies. To measure the correlation between proposed models and student performance, an experiment that involves 65 respondents was conducted. The data was analysed with structured and statistical approaches. Preliminary study shows that there are multivariate effects of the English fluency on PLA model. With the increasing demands of IT and software development skills, this research will help in motivating and encouraging more people to learn programming.

**Keywords** E-learning · Self-assessment · Programming · Learning assessment model

## 1 Introduction

In recent years, e-learning and online courses utilising electronic technologies represent a new learning mode for teaching a large number of students when face-to-face education is unlikely conducted. In contrast with the past, it's not hard to find an online platform that allows peer-to-peer communications and knowledge sharing in forms of readings, audio and video. When e-learning is grown, there is a need to address the real problem of helping programming novices to identify the difference between what they have expected to learn versus what they have actually achieved at certain points of time [1]. Without coaching in the physical classroom, students' failure in achieving positive learning outcomes might end with undesirable actions and behaviors. To promote student interest to learn and address student learning needs, self-assessment tool could be helpful with providing responsibly guidance to their direction and outcomes [2]. When the students are paid with continued close attention for their learning progress and gap, they can be given proper guidance in a timely manner [1]. Self-assessment has been proven that it is feasible to evaluate how the students are learning throughout a course with the existing set of materials when compared to summative assessment that evaluates how the students have learned at the end of the course [3].

## 2 Motivation

From the study, the knowledge background and skills of the students have impacts on their forthcoming trainings [4]. To reflex these factors when helping students understand the learning gaps and defects in their e-learning progress, **P**rogramming **L**earning **E**xperience **M**odel (PLEM) associated with **P**rogramming **L**earning **A**ssessment (PLA) activities, **A**utomated **A**ssessment (AA) and **L**anguage, **E**ducation, **A**chievement, and **P**rogramming (LEAP) attachment to measure student

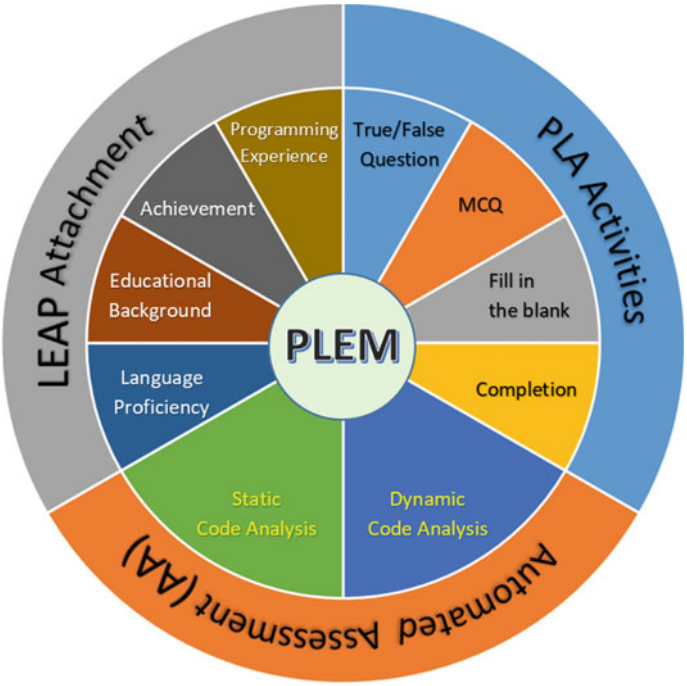



**Fig. 1** The components in the PLEM environment

performance and capability was proposed by Ho et al. [5]. Figure 1 shows the components in the PLEM environment. The PLA activities include four areas of *true/false question*, *multiple choice question (MCQ)*, *fill in the blank* and *completion*. The PLA activities help students obtain the necessary concepts to support them in designing solution for problem solving. The LEAP attachment highlights deep and enduring relationship between students' background and student performance levels.

It is believed that strong English language proficiency, strong programming experience and good achievement in previous examination are positively related to the students' performance in the PLA activities and automated assessment. But the education background of a student is negatively related to the mediation of instructor. Figure 2 shows the positive and negative correlations in the LEAP attachment.

## 3 Programming Learning Approach

E-learning has become a trend and essential in education. A general-purpose e-learning platform however, does not seem as useful but sometimes ineffective when

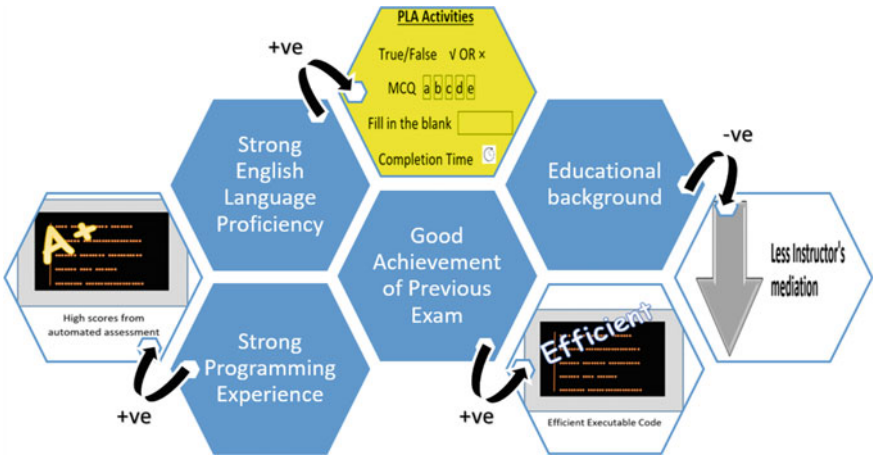



**Fig. 2** The positive and negative correlations in the LEAP attachment

it comes to programming education [2, 6]. Most of these platforms have assessment module in the system but methods and tools that provide scaling and personalised access for learning are deficient [4]. Figure 3 shows common practice of self-assessment in e-learning course.

To address aforementioned issues, we propose a personalised self-assessment access for the students after completing explicit LEAP attachment of the students and PLA activities. Students would get motivated to learn and explore the knowledge with an intelligent and personalised learning environment [7]. In the environment, involved parties, data and processes are to be taken into account as suggested by Bachir and Abenia [8]. Figure 4 shows the proposed work flow in PLEM. The actions highlighted in orange colour indicates the workflow for generating personalised self-assessment access. In the model, questions are categorised into concept-acquiring questions (highlighted in orange colour) and problem-solving questions (highlighted in green colour). Students get benefits if the necessary concepts obtained before they start to design solution for problem solving.

The process of evaluating code and generating useful feedback to the students may be cumbersome to the involved parties [9]. An automated assessment system to be designed so that programming teaching and learning could be supported. There are generally two approaches of automated assessment for programming. Static code analysis does not take program inputs into account [10]. It transforms the source-code into internal representations and compares them with the correct solutions in the repository [11]. Scores awarded for the matching attributes in the code. The following are examples of assessment system built using static code analysis:

- BOSS tests Java software automatically for comments in code, code style, code correctness and code structure. It mainly supports assessment process but not student learning [12].

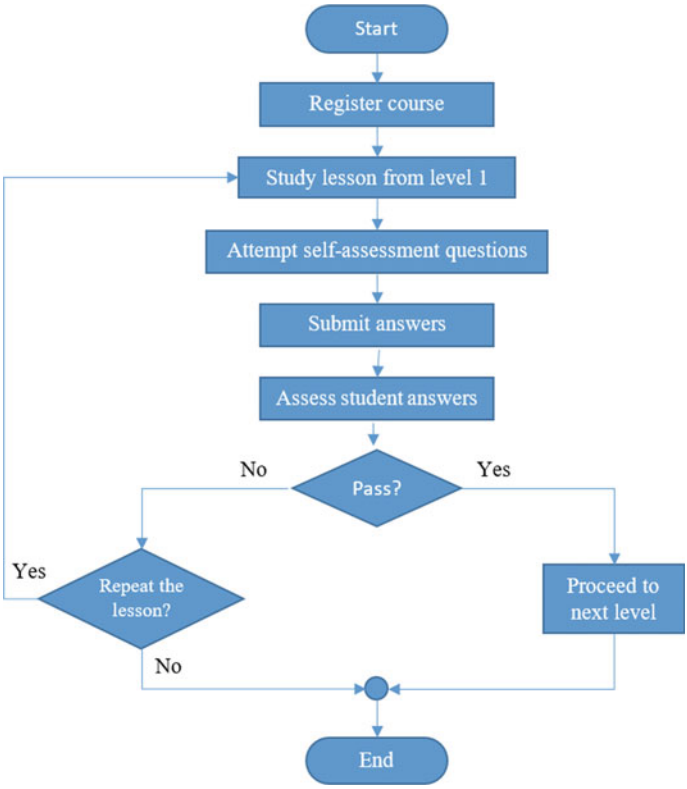



**Fig. 3** Common practice of self-assessment in e-learning course

- GradeIT applies program repair approach to provide feedback and grade for the submitted code. However, it's not able to assess complex code [13].
- Algo+, which is an assessment tool with referent solution in database for algorithmic competencies. Bey et al. aimed to study the quality of assessment but not learner's learning progress [11].
- UNED, which is an assessment system used by European distance universities. It supports huge number of student submissions and facilitates the process of giving feedback to students. The feedback is helpful in guiding students to complete their assignments. It reduces the number of teaching staff involved but human judgment is still required constantly in the practice [14].

On the other hand, dynamic code analysis takes program inputs into account. It is only able to work on compiled code. Scores awarded when the produced outputs match with the intended outputs. The examples of assessment system built using dynamic code analysis are EPFL grader which is an automated assessment tool developed at Swiss Federal Institute of Technology Lausanne [11], Mooshak which

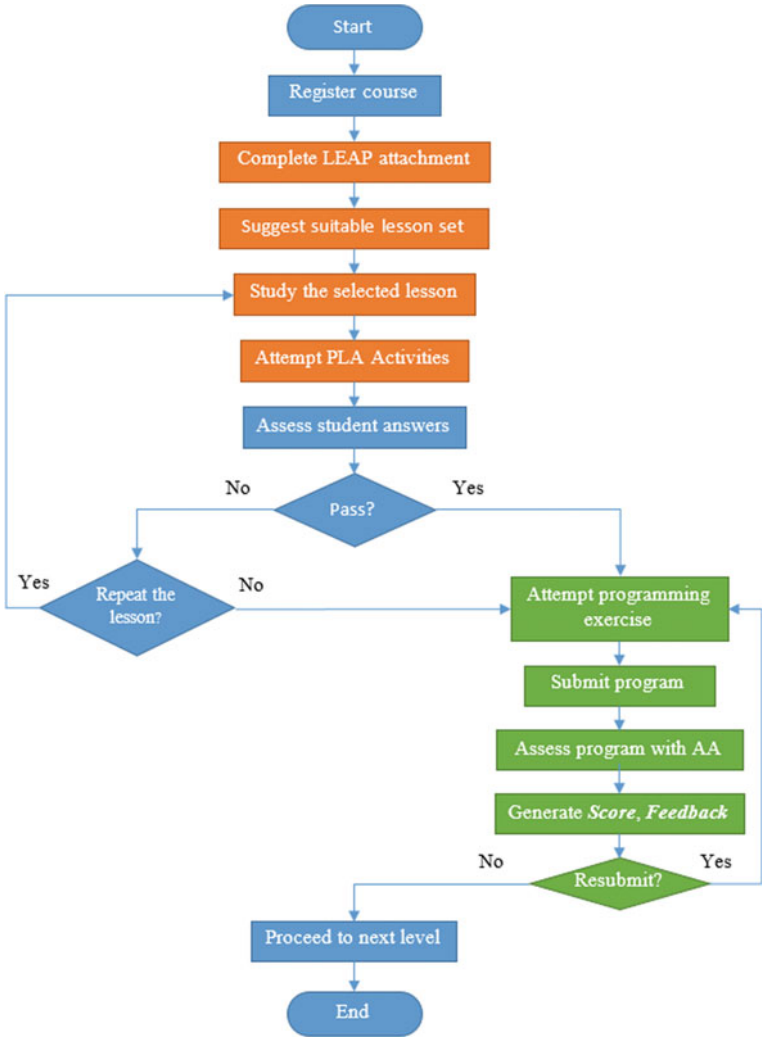



**Fig. 4** The proposed work flow in PLEM

is a Web-based automated assessment system developed for the online programming courses [12] and 2TSW which includes gamification features in the assessment system to motivate students in their learning [15]. Another tool is Dante used in TUL to validate high number of programming assignment submissions against different test cases. However, the involvement of lecturers is required for giving feedback to the works that cannot compile [16].

None of the above self-assessments take knowledge background of the students into account when design the system. With observation and exploration of the state of art of the above and other existing program assessment tools [17–20], a model for

students to experience better programming learning process to be designed. In the proposed model, LEAP attachment is incorporated with PLA activities to fit in with the students' needs.

## 4 Methodology

### *4.1 Experiment Design for PLA Model*

To prove the concepts highlighted in Motivation section, the programming self-assessment (or PLA—Programming Learning Assessment model) is systematically organised as the following:

- Quiz 1: Overview of C++ with fundamentals of programming
- Quiz 2: Functions with programming modules
- Quiz 3: Routines and control structures.

Here, we highlight guidelines for some of the measures observed throughout this research on the programming learning assessment (PLA) model.

(a) *Completion time estimation*: The cost of a project of learning technology includes the time and money expended in picking up new skills. The time factor is widely used in the assessment of programming tasks [3, 17]. First, the more data is collected, the more accurate will the estimate be. The major difficulty is that there may be new variables impacting upon the current task which have not been considered in the past. Second, a way to estimate completion time is through mathematical models.

(b) *Comprehension*: A high understanding score contributes to good students' perceptions of the given formative assessment [18–21], that is the ability to determine what a programming framework does and how it works, by going through the accompanying learning content. Comprehension has an inverse relation to complexity, as the complexity of a framework in question increases, the comprehension tends to decrease. This measure is assessed via the dimensions of True/False Questions, Multiple Choice Questions (MCQ), and Fill in the Blank questions.

(c) *Workings*: Parihar et al. proposed grading scores to obtain work repair feedback, which supports our usage of this dependent variable of workings [13]. This measure refers to the achievement requires specific knowledge of a particular topic as well as external factors such as the learning process and effort involved. We assess this measure using individual total score and time taken within Quiz 1, Quiz 2, and Quiz 3.

The sample size of these experimental subjects is 65, which is adequate since one needs at least 30 participants to be statistically valid. There are total 10 dependent variables (Y1 to Y10) from the PLA activities. The mapping of the dependent variables to the above measures as follows:

(a) Comprehension

- Y1-True/False (in summing the Q1Part1, Q2Part1, and Q3Part1 scores)
- Y2-MCQ (in combining Q1Part2, Q2Part2, Q3Part2 scores)
- Y3-Fill-in-the-blank (in summing Q1Part3, Q2Part3, and Q3Part3 scores)

(b) Y4-Completion time
(c) Workings

- Y5-Quiz1 total score, Y6-Quiz1 time taken
- Y7-Quiz2 total score, Y8-Quiz2 time taken
- Y9-Quiz3 total score, Y10-Quiz3 time taken.

## *4.2 Data Preprocessing for LEAP Model in Fluent English Language Proficiency*

To leverage the LEAP model, here is the intuition on English fluency based on the students' Achievement. Fluent English speaking student is denoted by $f$. Otherwise is denoted $n$ (non-fluent).

We analysed the ***English_Score*** attribute while looking at the LEAP attachment, i.e. **L**anguage Proficiency on English (Speaking, Listening, Reading, and Writing), **E**ducation background, the **A**chievement in ***English_Score***, and finally **P**rogramming Experience to determine the intuitions on English fluency as shown in Table 1.

The grouping of $f$ and $n$ are balanced across the 65 respondents. Table 2 further supports these two groups are balanced, where group $f$ does not score better than group $n$ in terms of their coursework mark (***cwMark***) achievement.

**Table 1** The intuitions on English fluency

| Intuition 1 | Intuition 2 | Intuition 3 | Intuition 4 | Grouping |
|---|---|---|---|---|
| *1. STPM GPA* | *2. UEC* | *3. MUET* | *4. Foundation English Grade* | *Fluency in English* |
| <3.3 | ≤B4 | Band 3 | <70 (B) | *n* (non-fluent) |
| ≥3.3 | B3<br>A2 | Band 4<br>Band 5 | 70–74 (B+)<br>75–79 (A−) | ***f* (Fluent English)** |

**Table 2** The means and standard deviations of the two groups

| LEAP model | *f* (Fluent English) | *N* (non-fluent) |
|---|---|---|
| N (participants) | 32 | 33 |
| Mean (cwMark) | 69.53 | 69.52 |
| Std. dev. (cwMark) | 15.713 | 11.333 |

**Table 3** Results of normality test

| Category (Dependent variable) | Kolmogorov-Smirnov | Category (Dependent variable) | Kolmogorov-Smirnov |
|---|---|---|---|
| | Statistic | | Statistic |
| 1. True/False question | 0.847 | 6. Quiz1 time taken | 1.304 |
| 2. MCQ | 1.011 | 7. Quiz2 total score | 1.320 |
| 3. Fill-in-the-blank | 1.050 | 8. Quiz2 time taken | 1.103 |
| 4. Completion time | 0.778 | **9. Quiz3 total score** | **1.618**** |
| 5. Quiz1 total score | 0.971 | 10. Quiz3 time taken | 1.347 |

**Significant at 0.05 level

## *4.3 Results and Analysis*

From the result of normality test of 1-Sampe K-S (Kolmogorov-Smimov) in Table 3, all the above dependent variables are normally distributed data, except Y9-Quiz3 total score. Y9-Quiz3 total score, which has Asymptotic Significance (2-tailed) [*p*-value] < 0.05, indicating this variable is not normally distributed, and needs non-parametric test, such as Mann Whitney u test. For normally distributed data, we can adopt parametric test such as MANOVA (Multivariate Analysis of Variance) subsequently.

Table 4 shows the fluent English group has consistently higher average of scores, and longer time taken across Y4 Completion time, Y6-Quiz1 time taken, Y8-Quiz2 time taken, and Y10-Quiz3 time taken. Fluent English speaking students perform better in programming than non-fluent ones. Fluent ones spend longer time in answering the exercise carefully.

Table 5 shows the results using Wilks' Lambda = 0.803, $F(8, 56) = 1.723$ ($p > 0.113$) indicated no evidence of significant difference among the mean scores.

Next, we highlight the contribution of the work here by elaborating the analysis of results collected from 65 respondents. The English score obtained from the LEAP model helps us to map the *Excellent* (*E*) and *Good* (*G*) groups to 32 fluent English speaking students. Meanwhile, the remaining 33 non-fluent students are mapped to the *Moderate* (*M)* and *Poor* (*P*) groups. Among the dependent variables within *Comprehension* from the PLA model analysed through parametric test of MANOVA (Multivariate Analysis of Variance) in Table 6, we see that the treatments in Y1-True/False Question has a strong significant difference at the 0.050 level (significance or *p*-value < 0.050). Y3-Fill-in-the-blank and Y7-Quiz2 total score have marginal

**Table 4** The means, median and standard deviations of all categories

| Category (Dependent variable) | Mean | | Std.dev. | |
|---|---|---|---|---|
| | *f* | *n* | *f* | *N* |
| 1. True/False question (scale: 0–6) | 20.69 | 17.09 | 4.388 | 6.943 |
| 2. MCQ (scale: 0–15) | 9.53 | 8.48 | 3.538 | 3.667 |
| 3. Fill-in-the-blank (scale: 0–15) | 10.38 | 8.61 | 3.536 | 4.486 |
| 4. Completion time (mm:ss) | 34:26 | 30:30 | 15:06 | 17:20 |
| 5. Quiz1 total score (scale: 0–20) | 12.63 | 11.09 | 3.643 | 4.156 |
| 6. Quiz1 time taken (mm:ss) | 14:35 | 11:14 | 9:46 | 6:56 |
| 7. Quiz2 total score (scale: 0–20) | 15.00 | 12.33 | 5.035 | 5.998 |
| 8. Quiz2 time taken (mm:ss) | 9:28 | 9:12 | 5:47 | 6:50 |
| 9. Quiz3 time taken (mm:ss) | 10:21 | 10:03 | 7:47 | 8:54 |
| 10. Quiz3 total score (scale: 0–20) | 16.50 | 13.00 | 7.403 | 7.229 |

**Table 5** Wilks' lambda and F test for the multivariate effect (English fluency)

| Wilks' lambda | Value | Exact F | Significance |
|---|---|---|---|
| df = (8, 56) | 0.803 | 1.723 | 0.113 |

**Table 6** Multivariate effects of the English fluency on PLA model, three specific scores, and time taken of quizzes

| No. | Categories | *F* | $Eta^2$ | *p*-value | No. | Categories | *F* | $Eta^2$ | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| 1. | True/False question | **6.188**** | **0.089** | **0.016**** | 6. | Quiz1 time taken | 2.558 | 0.039 | 0.115 |
| 2. | MCQ | 1.370 | 0.021 | 0.246 | 7. | Quiz2 total score | **3.757*** | **0.056** | **0.057*** |
| 3. | Fill-in-the-blank | **3.105*** | **0.047** | **0.083*** | 8. | Quiz2 time taken | 0.031 | 0.000 | 0.861 |
| 4. | Completion time | 0.951 | 0.015 | 0.333 | 9. | Quiz3 time taken | 0.021 | 0.000 | 0.884 |
| 5. | Quiz1 total score | 2.498 | 0.038 | 0.119 | | | | | |

**$p < 0.050$ (2-tailed); *$p < 0.100$ (2-tailed)

differences at the 0.100 level (*p*-value < 0.100). There is no significant difference between the groups in Y2-MCQ, Y4-*Completion time*, and other variables. There is an interesting finding here. Fluency in English helps students to perform significantly better in True/False Question and Fill in the Blank nature. If one does not want the influence of English fluency affecting the students' performance in introductory programming, the nature of questions can be formatted in multiple choice questions.

Finally, the non-parametric test of Mann Whitney in Table 7 indicates that Y9-Quiz3 total score make marginally difference at the 0.100 (*p*-value < 0.100). The

**Table 7** Mann-Whitney test results on the number of difficulties

| **Group** | **Sample size, n** | **Mean rank** | **Sum of ranks** | **Removal of invalid cases** |
|---|---|---|---|---|
| 1. Fluent in English | 32 | 37.27 | 1192.50 | No invalid case |
| 2. Non-fluent | 33 | 28.86 | 952.50 | No invalid case |
| **Test Statistics** | **Mann-Whitney U** | **Z** | **Wilcoxon W** | **Asymptotic Sig. (2-tailed)** |
| Quiz3 total score | 391.500 | −1.810 | 952.500 | **0.070*** |

*$p < 0.100$ (2-tailed)

results support that English fluency does play an influence on learning programming, especially when the student gets to more advanced topics such as functions and recursion, as compared to the overview of C++.

## 5 Justification

In a recent systematic review on programming pedagogy by Medeiros et al., inadequate background knowledge in students to learn programming was highlighted [4]. When come to the methods to improve the capabilities of students who do not have adequate background knowledge, it is necessary to establish metrics to evaluate the effectiveness [4]. The proposed LEAP and PLA models can fill in the gap between the background knowledge and student competencies. The cognitive and language development in learners is also very important to learn programming [21]. Our findings further affirm that subjects with English proficiency performed significantly better than non-fluent subjects in True/False questions and Fill in the Blank questions. The PLA model also helped the fluent ones (similar to *E* and *G* groups) master the more advanced topics such those found in programming course, as compared to another non-fluent group. The findings can be referenced as the base in developing an intelligent and personalised learning environment—PLEM. We believe that when all learning aspects are integrated and managed well, a higher level of student learning experience can be achieved.

**Acknowledgements** The authors would like to thank the financial support given by the Fundamental Research Grant Scheme, FRGS/1/2019/SS06/MMU/02/4 and Multimedia University, Cyberjaya, Malaysia (Project ID: MMUE/190031).

## References

1. Krusche S, Seitz A (2018) ArTEMiS—an automatic assessment management system for interactive learning. In: Proceedings of the 49th ACM technical symposium on computer science

education (SIGCSE '18), Baltimore, MD, USA. ACM, New York, NY, pp 284–289
2. Shohel Rana M, Bhuiyan T, Satter AKMZ (2018) e-school: design and implementation of web based teaching institution for enhancing E-learning experiences. In: Nguyen N, Pimenidis E, Khan Z, Trawiński B (eds) Computational collective intelligence. ICCCI 2018. Lecture notes in computer science, vol 11055. Springer, Cham
3. Pieterse V, Liebenberg J (2017) Automatic vs manual assessment of programming task. In: Proceedings of the 17th Koli calling international conference on computing education research (Koli calling '17), Koli, Finland. ACM, New York, NY, pp 193–194
4. Medeiros RP, Ramalho GL, Falcão TP (2019) A systematic literature review on teaching and learning introductory programming in higher education. IEEE Trans Educ 62(2):77–90
5. Ho SB, Chean SL, Chai I, Tan CH (2019) Engineering meaningful computing education: programming learning experience model. In: Proceedings of 2019 IEEE international conference on industrial engineering and engineering management (IEEM), 15–18 Dec 2019, Macao. IEEE, pp 925–929
6. Alzaid M, Trivedi D, Hsiao I (2017) The effects of bite-size distributed practices for programming novices. In: 2017 IEEE frontiers in education conference (FIE), Indianapolis, IN, pp 1–9. https://doi.org/10.1109/fie.2017.8190593
7. Muñoz García A, Lamolle M, Martinez-Béjar R, Espinal Santana A (2019) Learning ecosystem ontology with knowledge management as a service. In: Nguyen N, Chbeir R, Exposito E, Aniorté P, Trawiński B (eds) Computational collective intelligence. ICCCI 2019. Lecture notes in computer science, vol 11684. Springer, Cham
8. Bachir S, Abenia A (2019) Internet of everything and educational cyber physical systems for university 4.0. In: Nguyen N, Chbeir R, Exposito E, Aniorté P, Trawiński B (eds) Computational collective intelligence. ICCCI 2019. Lecture notes in computer science, vol 11684. Springer, Cham
9. Chen LS, Chen CC, Chang SH, Yang E (2017) An e-learning system for programming languages with semi-automatic grading. In: 10th international conference on ubi-media computing and workshops, Pattaya, Thailand, pp 356–361
10. Louridas P (2006) Static code analysis. IEEE Softw 23(4):58–61. https://doi.org/10.1109/MS.2006.114
11. Bey A, Jermann P, Dillenbourg P (2018) A comparison between two automatic assessment approaches for programming: an empirical study on MOOCs. Educ Technol Soc 21(2):259–272
12. Fernández-Alemán JL (2011) Automated assessment in a programming tools course. IEEE Trans Educ 54(4):576–581
13. Parihar S, Dadachanji Z, Singh PK, Das R, Karkare A, Bhattacharya A (2017) Automatic grading and feedback using program repair for introductory programming courses. In: Proceedings of 22nd annual conference on innovation and technology in computer science education (ITiCSE'17), Bologna, Italy. ACM, New York, NY, pp 92–97
14. Galan D, Heradio R, Vargas H, Abad I, Cerrada JA (2019) Automated assessment of computer programming practices: the 8-years UNED experience. IEEE Access 7:130113–130119. http://doi.org/10.1109/ACCESS.2019.2938391
15. Polito G, Temperini M, Sterbini A (2019) 2TSW: automated assessment of computer programming assignments, in a gamified web based system. In: 18th international conference on information technology based higher education and training (ITHET), Magdeburg, Germany, pp 1–9. http://doi.org/10.1109/ITHET46829.2019.8937377
16. Duch P, Jaworski T (2018) Dante—automated assessments tool for students' programming assignments. In: 11th international conference on human system interaction (HSI), Gdansk, pp 162–168. http://doi.org/10.1109/HSI.2018.8431146
17. Tek FB, Benli KS, Deveci E (2018) Implicit theories and self-efficacy in an introductory programming course. IEEE Trans Educ 61(3):218–225
18. Ogange BO, Agak JO, Okelo KO, Kiprotich P (2018) Student perceptions of the effectiveness of formative assessment in an online learning environment. Open Prax 10(1):29–39
19. Pyper A (2018) Student perceptions of the implementation of formative assessment: a Royal St. George's college case study. Young Researcher 2(1):135–147

20. Voinea L (2018) Formative assessment as assessment for learning development. J Pedagogy 1:7–23
21. Strawhacker A, Bers MU (2019) What they learn when they learn coding: investigating cognitive domains and computer programming knowledge in young children. Educ Technol Res Dev 67(3):541–575. http://doi.org/10.1007/s11423-018-9622-x

# Design and Functionality of a University Academic Advisor Chatbot as an Early Intervention to Improve Students' Academic Performance


**Mei Shyan Lim, Sin-Ban Ho, and Ian Chai**

**Abstract** This paper introduces the design and functionality of a university academic advisor chatbot, which leverages on the result of a prediction model to predict students' academic performance, to do early intervention to assist students who may need academic guidance. The prediction model is based on students' attendance and scores of formative assessments to predict the score of the final summative assessment using a suitable machine learning algorithm. Scikit-learn library using Python will be used in this research to run the machine learning algorithms. The chatbot will be developed using Dialogflow which is integrated with one of the text messaging apps and established connection to a database. The database stores students' attendance, scores of formative assessments, scores of final summative assessments and the status of students whom the chatbot has reached out to. This research aims to reduce the workload of lecturers to reach out to every student who is predicted to have problems in their academic studies and at the same time, be able to assist students using a chatbot.

**Keywords** Learning analytics · Machine learning · Chatbot · Prediction · Tertiary education

M. S. Lim (✉) · S.-B. Ho · I. Chai
Faculty of Computing and Informatics, Multimedia University, 63100, Cyberjaya, Selangor, Malaysia
e-mail: 1191400099@student.mmu.edu.my

S.-B. Ho
e-mail: sbho@mmu.edu.my

I. Chai
e-mail: ianchai@mmu.edu.my

M. S. Lim
Faculty of Computing and Information Technology, Tunku Abdul Rahman University College, 53300 Kuala Lumpur, Wilayah Persekutuan, Malaysia



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_15

## 1 Introduction

Tertiary education plays a very important role for the development of a nation and the catalyst for the economic growth. Generally, the vision and mission of Institutions of Higher Learning (IHLs) have always focus on holistic education to work towards students' success. However, it is observed that many IHLs are facing with increasingly high students' dropout, mainly due to lack of motivation or interest to continue with their studies [1].

Therefore, giving timely support to students who are having problems in their studies is always the effort made by education providers to increase the percentage of students' success [2] and increase graduation rates [3]. Currently, it is difficult to depend solely on lecturers to identify every individual student facing problems at the early stage of the weeks for them to do intervention, especially if dealing with large group of students. To address this issue, there are quite a number of research which used machine learning algorithms and learning analytics to identify students who may have problems to cope with their studies during early weeks of their studies [2, 4].

Learning Analytics is the process of collecting, measuring and analysing data about students as learners with the objective to improve the quality of learning, which can be used to provide performance feedback and learning recommendation to students with different learning behavior [5].

Tempelaar et al. [6] stated that feedback to students is only helpful if two conditions are met:

- it is predictive
- intervention can be made on time to support students.

For this proposed research, a prediction model will be developed to predict students' performance of a course, based on past data. It is expected that an early intervention can be taken to assist and guide a student whom is predicted not able to cope with a course that he or she is undergoing.

Since there were so many studies done to investigate which machine learning algorithm provides the most accurate prediction in the aspect of student retention, the focus of this research is to utilize the result of the best machine learning algorithm used to trigger a chatbot via a text messaging app to complement the educators' role to assist students at the preliminary stage. The study will be tested for students who are undergoing tertiary education.

Although a common challenge faced by researchers in this area is to have a generic prediction model to fit for different courses in different universities [4, 7], solving this problem will not be the focus of this research, but to be considered to be solved in future research.

## 2 Related Work

Literature review about the prediction model used to predict students' motivation or achievement in their studies, followed by usage of chatbot in education will be conducted.

Various researches have been conducted to predict students' motivation or achievement via distance learning system, Massive Open Online Courses (MOOCs), blended learning or learning management systems in different level of studies which is K12 or Tertiary Education. Learning Analytics (LA) provides useful insights to students, educators or institutions of higher learning (IHLs) in predicting students' success by leveraging on Machine Learning (ML) techniques using large amount of educational data collected [8].

Random Tree algorithm which is one of the Tree Classifier algorithms in WEKA has the best accuracy to predict and classify academic performance [9]. In another piece of research, the decision Tree J48 model showed the best result to identify students' low-engagement based on the data collected from an e-learning system. The most significant input variables were clicks on study materials, discussion forums, students' navigation path and homepage. However, it was stated that student engagement is a very complex problem and may depend on other factors like teaching experience, course design, teaching style and course concepts [10].

A predictive model using machine learning tools and techniques are used to identify students as early as Week 3 of a Fall semester for early intervention to improve student retention. A combination of classifiers (Random Tree is tested the best classifier) and measurements (mainly students' attendance) have been selected and achieved in excess of 97% of accuracy. However, it is stated that machine learning applications will not remain static as training the machine learning model is an ongoing process to include new data and to remove aging data. There is also a need to do more research across other institutions of higher learning with similar characteristics of students to validate whether the predictive model still produce accurate results [4]. Table 1 shows the summary of related work about machine learning algorithms predicting students' academic achievement, graduation time, drop-out and motivation.

In summary, educational data that can be collected comprises of the individual or combination of the followings to do prediction [14]:

- Students' Attendance [4]
- Students' Academic Performance in summative or formative assessment [9, 12]
- Students' Interaction within LMS [9, 12, 15, 16]
- Students' Personality [17]
- Students' Motivation [9, 17–19]
- Students' Demographic Profile [11].

In terms of the tools used that are related to learning analytics and machine learning, WEKA is the popular open source machine learning toolkit used in many pieces of research [4, 9, 10]. Another very popular open source machine learning

**Table 1** Summary of related works

| Selected predictors variables | Machine learning algorithms | Prediction | Authors |
|---|---|---|---|
| Students' attendance | Random tree | Students' grade | Gray and Perkins [4] |
| • Velocity—speed<br>• Quantity—in terms of length of answers<br>• Relevancy—correctness of answers | Random tree | Students' motivation | Juliane et al. [9] |
| • Student demographics profile (place of birth, age)<br>• Organization (activist, non-activist) | C4.5 decision tree | Students' graduation time | Budiman et al. [11] |
| 1. Persistence<br>• Total number of visits to learning management system (LMS)<br>• Duration of watching video<br>• Completion of tasks<br>2. Enthusiasm<br>• Post and reply in forum<br>3. Responsibility<br>• Average score of all homework<br>• Average score of scores of the 10 examinations<br>• Average score of submission timestamps of the 10 examinations | Deep neural network (DNN) | Students' grades | Yu [12] |
| • Sociodemographic<br>• Program<br>• Academic History | Random forest | Students' dropout | Solis [13] |

toolkits on the rise are Python, Scikit-Learn machine learning library [20–25] and RapidMiner [11, 15] which are very promising to be applied in both research and real-life applications.

Once a student who is undergoing a course through an e-learning platform is predicted to have loss of motivation or poor academic performance, Artificial Intelligence (AI) can be applied to construct adaptive learning mechanisms using recommender systems [26]. However, it was found that AI is commonly known as a complex solution that is more expensive, prone to error if it is not designed accurately and requires a higher level of technical skills to implement. All risks and obstacles should be properly addressed to apply AI in e-learning.

Application of AI through chatbot is on the rise to solve many real-world applications. Chatbot is a conversational agent that interacts with human being with natural language processing capability [27]. Although usage of chatbot in education is still at

infancy stage, there were numerous positive outcomes based on the research done as of now. It was tested on various areas in education sector, for example as Intelligent Tutoring System [28], to support Admission Services [29], to provide Academic Advisement [30], to support first year students [31] and to support General Student Enquiries [32–34]. Most of the chatbots focus on using text-based messaging. Recent research work for chatbot is related to voice recognition, speech-to-text capability [29] and adaptive learning [28] to name a few.

In the aspect of chatbot engine, recent research work focused in improving knowledge-based response using rule-based approach [29, 32], Artificial Intelligence Markup Language (AIML) [33], message preprocessing approach [34], answer from Question and Answer Database [35] and Natural Language Processing (NLP) [36].

Based on the literature review, there is a research gap in using a chatbot as an early intervention to assist university students' who are predicted to have poor academic performance using suitable machine learning algorithms. The focus of the research is not to develop a comprehensive knowledge-based chatbot but to leverage on chatbot using one of the social messaging app to remind the students to attend classes or to alert them to do mini-revision and supplementary quizzes as part of the early intervention process.

## 3 Methodology

The proposed solution to increase the students' performance in tertiary education is to implement early intervention during the early weeks of a semester by using a predictive model to trigger a University Academic Advisor chatbot using one of the text messaging platforms (e.g. Whatsapp,[1] Telegram,[2] Hangouts Chat[3]) to reach out to students who may potentially face problems in their studies.

Empirical study will be carried out for this research. Datasets will be collected from one common course offered by a university for two semesters. It is estimated to involve 100–150 students. Comparison for both experimental and controlled groups will be made to analyse whether there is any significant impact of a chatbot as an early intervention to improve students' academic performance. Figure 1 shows the overview of the methodology.

### *3.1 Data Collection and Preparation*

Two datasets will be collected for a course which is offered in two different semesters. The first dataset is used as the training and validation purpose and the second dataset

[1] https://www.whatsapp.com/.

[2] https://telegram.org/.

[3] https://hangouts.google.com/.

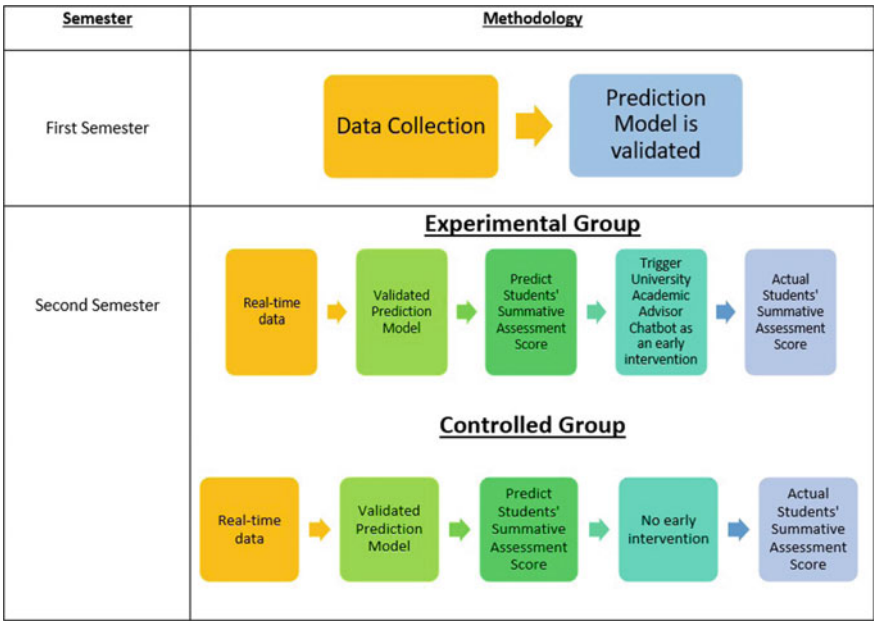



**Fig. 1** Overview of the methodology

is used to predict who are the students predicted to have low summative assessment score and to reach out to these students by using the university academic advisor chatbot.

The features of the dataset are:

- **Att_1, Att_2…Att_n** represents the total number of attendances for tutorial and/or practical class on a weekly basis where *n* represents the total number of weeks
- **FA_1**, **FA_2…FA_n** represents the score of every formative assessment where *n* denotes the total number of formative assessments conducted throughout the semester.
- **Final SA Score** represents the score of a final summative assessment.

## *3.2 Feature Selection and Machine Learning Algorithms*

Feature selection will be made to provide the best fit for the prediction model using a suitable classification machine learning algorithm which is to predict whether a student will pass or fail the final summative assessment of a course.

Tools to be used for feature selection, learning algorithms and validation of models are Python using Scikit-Learn[4] library which consists of various machine learning algorithms.

### 3.3 Preliminary Data Analysis

Preliminary testing was conducted using the Educational Processing Mining dataset which is a public dataset obtained from the UCI Machine Learning Repository website [37]. The dataset, which closely matched with the features of the proposed dataset, recorded the sessions attended by the 115 first year undergraduate students, major in Engineering at the University of Genoa together with their formative and summative assessment marks. There were 6 sessions of participation, formative assessment for Session 2 to Session 6 and a final summative assessment. Students' participation for every session closely matched with students' attendance.

Random forest model is selected for the preliminary testing based on its popularity as a supervised machine learning algorithm with high prediction accuracy and with feature selection capability [38, 39]. Since the goal of this research is early intervention, data obtained during early weeks are selected as possible features for the prediction model. Training dataset and testing dataset are split to 80% and 20% respectively. Using the feature importance, it was found that the grade of Session 2 (*S2_FA*) and Session 3 (*S3_FA*) formative assessment made significant impact to predict students' academic performance (pass or fail) as compared to attending respective sessions (*S1*, *S2*, *S3*) as shown in Fig. 2.

Table 2 showed that the highest accuracy to predict students' academic performance using the Educational Processing Mining public dataset is to train the model using 90% of the data. Although the F1-Score (Weighted Average) is 0.75, this is a good indication of students possibly having problems at the early stage of the studies and to provide early intervention to assist them.

### 3.4 Design and Functionality of the University Academic Advisor Chatbot

The University Academic Advisor Chatbot will be developed using Dialogflow.[5] The chatbot will be deployed using one text messaging app. A database is used to store the students' attendance, scores of formative assessments, scores of final summative assessments and the status of students whom the chatbot has reached out to for subsequent follow-up. The chatbot will take necessary actions as shown in Table 3. Sequence diagram of the University Academic Advisor Chatbot is depicted in Fig. 3.

[4] https://scikit-learn.org/.

[5] https://dialogflow.com/.

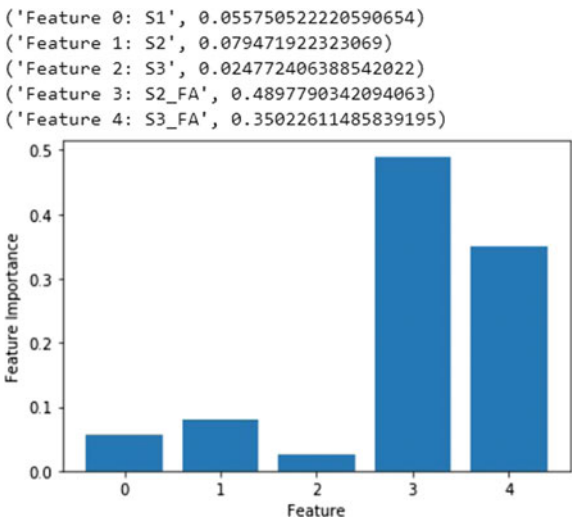



**Fig. 2** Feature importance for the participation of Session 1 (*S1*), Session 2 (*S2*), Session 3 (*S3*), formative assessment for Session 2 (*S2_FA*) and Session 3 (*S3_FA*)

**Table 2** Comparison of F1-score (weighted average) according to the percentage of training data and testing data

| Training data (%):Testing data (%) | F1-score (weighted average) |
|---|---|
| 70:30 | 0.65 |
| 75:25 | 0.61 |
| 80:20 | 0.65 |
| 85:15 | 0.72 |
| 90:10 | 0.75 |

## 4 Conclusion

Maintaining students' retention in Institutions of Higher Learning is always a big challenge. There are many researches focusing in finding the most suitable machine learning algorithms to predict students' academic performance, motivation and drop-out with high accuracy but there are not many research in utilizing chatbot as an early intervention to assist students who may need lecturers' help at university level.

The scope of this research is to:

- choose most suitable machine learning technique(s) using new datasets (attendance and scores of formative assessment) to accurately predict the students' score for the final summative assessment.

**Table 3** Actions by the university academic advisor chatbot

| Condition | Actions by chatbot |
|---|---|
| Attendance is poor | Remind and ask whether the student will attend the next classes<br>If the answer is yes, provide a thank you message to the student for confirming their attendance. At the same time, provide encouragement or inspiring words to further motivate the student to attend class<br>If the answer is no, list all possible reasons the student is not able to attend class for the student to choose from. Chatbot will provide pre-defined solutions for the reason that student cannot attend the next class<br>If there is no reply from the student within 24 h, the chatbot will remind and ask again whether the student will attend the next class<br>If there is no reply from the student within 48 h or the reply is not able to attend the next class, the chatbot will send an alert to the lecturer in charge of the course for human intervention |
| Average score of identified learning activities is low | Based on which topic that the student has demonstrated low achievement based on the score of the identified learning activities, chatbot will provide mini revision, by prompting the student to access to the learning materials and do a supplementary quiz to strengthen the foundation<br>If there is no reply from the student within 24 h, the chatbot will attempt to prompt the student to access to the learning materials and do a supplementary quiz to strengthen the foundation<br>If there is no reply from the student within 48 h or still do not do well in the supplementary quiz, the chatbot will send an alert to the lecturer in charge of the course for human intervention |

- develop a university academic advisor chatbot using a text messaging app as the platform to reach out to students to address students' absenteeism or low average score for formative assessment.
- send alert messages to lecturers to conduct human intervention, if students do not have any response after two attempts made by the university academic advisor chatbot.

Using machine learning algorithm and the Educational Processing Mining public dataset has provided some insights of the impact of some features as compared to the other non-significant features to predict students' academic performance which will be further analyzed using new datasets to be collected soon.

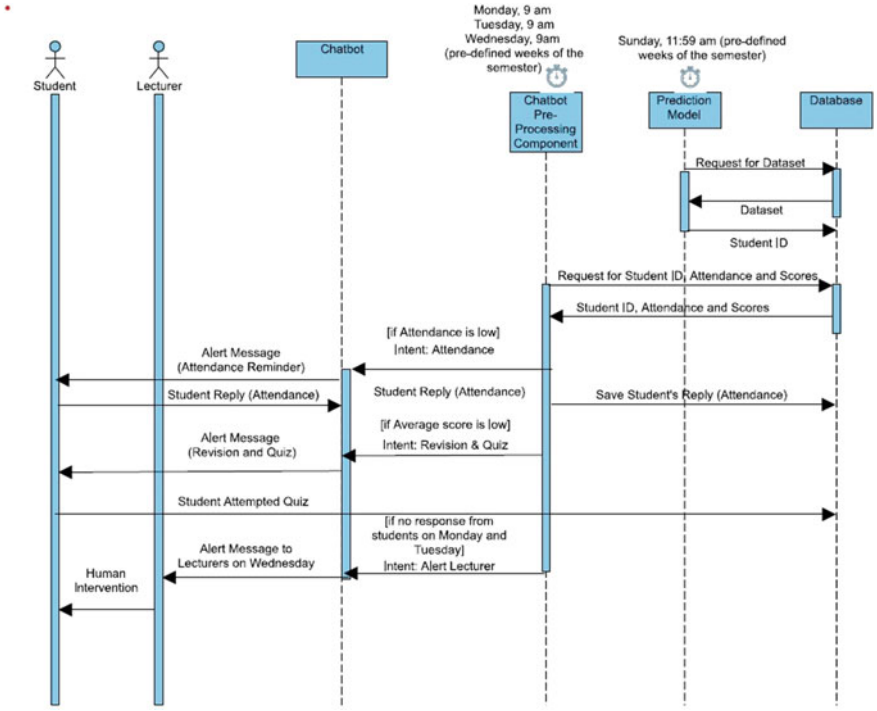



**Fig. 3** Sequence diagram of the university academic advisor chatbot

## References


1. Pellegrini G, Segafredo C (2015) Keeping pace: educational choice motivations and first-year experiences in the words of Italian students. In: Understanding student participation and choice in science and technology education. Springer, Netherlands, pp 259–273. https://doi.org/10.1007/978-94-007-7793-4_16
2. Shehata S, Arnold KE (2015) Measuring student success using predictive engine. In: Proceedings of the fifth international conference on learning analytics and knowledge—LAK 2015. ACM Press, New York, pp 416–417. https://doi.org/10.1145/2723576.2723661
3. Sudipa IGI, Wijaya INSW, Radhitya ML, Mahawan IMA, Arsana INA (2020) An android-based application to predict student with extraordinary academic achievement. In: journal of physics: conference series, vol 1469. Institute of Physics Publishing, Bali. https://doi.org/10.1088/1742-6596/1469/1/012043
4. Gray CC, Perkins D (2019) Utilizing early engagement and machine learning to predict student outcomes. Comput Educ 131:22–32. https://doi.org/10.1016/j.compedu.2018.12.006
5. Keller B, Baleis J, Starke C, Marcinkowski F (2019) Machine learning and artificial intelligence in higher education: a state-of-the-art report on the German University landscape. https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft/KMW_I/Working_Paper/Keller_et_al.__2019__-_AI_in_Higher_Education.pdf
6. Tempelaar DT, Rienties B, Giesbers B (2015) In search for the most informative data for feedback generation: learning analytics in a data-rich context. Comput Hum Behav 47:157–167. https://doi.org/10.1016/j.chb.2014.05.038

7. Sohail S, Khanum A, Alvi A (2019) Hybrid fuzzy-statistical system for learning analytics. In: Proceedings of 2018 IEEE international conference on teaching, assessment, and learning for engineering, TALE 2018, pp 989–994. https://doi.org/10.1109/TALE.2018.8615182
8. Gkontzis AF, Kotsiantis S, Tsoni R, Verykios VS (2018) An effective LA approach to predict student achievement. In: ACM international conference proceeding series, pp 76–81. https://doi.org/10.1145/3291533.3291551
9. Juliane C, Arman AA, Sastramihardja HS, Supriana I (2017) Predicting the presence of learning motivation in electronic learning: a new rules to predict. TELKOMNIKA (Telecommun Comput Electron Control) 15(3):1223–1229. https://doi.org/10.12928/telkomnika.v15i3.4286
10. Hussain M, Zhu W, Zhang W, Abidi SMR (2018) Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput Intell Neurosci 2018:1–21. https://doi.org/10.1155/2018/6347186
11. Budiman E et al (2018) Performance of decision tree C4.5 algorithm in student academic evaluation. In: Alfred R, Iida H, Ibrahim A, Lim Y (eds) Computational science and technology. ICCST 2017. Lecture notes in electrical engineering, vol 488. Springer, Singapore, pp 380–389. https://doi.org/10.1007/978-981-10-8276-4_36
12. Yu C (2018) SPOC-MFLP: a multi-feature learning prediction model for SPOC students using machine learning. J Appl Sci Eng 21(2):279–290. https://doi.org/10.6180/jase.201806_21(2).0016
13. Solis M, Moreira T, Gonzalez R, Fernandez T, Hernandez M (2018) Perspectives to predict dropout in university students with machine learning. In: 2018 IEEE international work conference on bioinspired intelligence (IWOBI). IEEE. https://doi.org/10.1109/IWOBI.2018.8464191
14. Ortigosa A et al (2019) From lab to production: lessons learnt and real-life challenges of an early student-dropout prevention system. IEEE Trans Learn Technol 12(2):264–277. https://doi.org/10.1109/TLT.2019.2911608
15. Almeda MV et al (2018) Comparing the factors that predict completion and grades among for-credit and open/MOOC students in online learning. Online Learn J 22(1):1–18. https://doi.org/10.24059/olj.v22i1.1060
16. Abou Gamie E, Abou El-Seoud S, Salama M, Hussein W (2019) Multi-dimensional analysis to predict students' grades in higher education. Int J Emerg Technol Learn (iJET) 14(02):4. https://doi.org/10.3991/ijet.v14i02.9905
17. Kubiatko M, Hsieh M-Y, Ersozlu ZN, Usak M (2018) The motivation toward learning among Czech high school students and influence of selected variables on motivation. Revista de Cercetare si Interventie Sociala 60:79–93
18. Al-Shabandar R, Hussain A, Laws A, Keight R, Lunn J (2017) Towards the differentiation of initial and final retention in massive open online courses. In: 13th international conference on intelligent computing, ICIC 2017, vol 10361. Springer, Liverpool, United Kingdom, pp 26–36. https://doi.org/10.1007/978-3-319-63309-1_3
19. Schumacher C, Ifenthaler D (2018) The importance of students' motivational dispositions for designing learning analytics. J Comput High Educ 30(3):599–619. https://doi.org/10.1007/s12528-018-9188-y
20. Amigud A, Arnedo-Moreno J, Daradoumis T, Guerrero-Roldan A-E (2017) Using learning analytics for preserving academic integrity. Int Rev Res Open Distrib Learn 18(5):192–210. https://doi.org/10.19173/irrodl.v18i5.3103
21. Azcona D, Hsiao I-H, Smeaton AF (2019) Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. User Model User-Adap Inter 29(4):759–788. https://doi.org/10.1007/s11257-019-09234-7
22. Gitinabard N, Xu Y, Heckman S, Barnes T, Lynch CF (2019) How widely can prediction models be generalized? Performance prediction in blended courses. IEEE Trans Learn Technol 12(2):184–197. https://doi.org/10.1109/TLT.2019.2911832
23. Hew KF, Qiao C, Tang Y (2018) Understanding student engagement in large-scale open online courses: a machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs. Int Rev Res Open Distrib Learn 19(3):69–93. https://doi.org/10.19173/irrodl.v19i3.3596

24. Hlosta M, Zdrahal Z, Zendulka J (2017) Ouroboros. In: Proceedings of the seventh international learning analytics & knowledge conference. ACM, Vancouver, pp 6–15. https://doi.org/10.1145/3027385.3027449
25. Doleck T, Lemay DJ, Basnet RB, Bazelais P (2020) Predictive analytics in education: a comparison of deep learning frameworks. Educ Inf Technol 25(3):1951–1963. https://doi.org/10.1007/s10639-019-10068-4
26. Wan S, Niu Z (2020) A hybrid e-learning recommendation approach based on learners' influence propagation. IEEE Trans Knowl Data Eng 32(5):827–840. https://doi.org/10.1109/TKDE.2019.2895033
27. Shawar BA, Atwell E (2015) ALICE chatbot: trials and outputs. Computación y Sistemas 19(4):625–632. https://doi.org/10.13053/cys-19-4-2326
28. Fadhil A, Villafiorita A (2017) An adaptive learning with gamification & conversational UIs: the rise of CiboPoliBot. In: Adjunct publication of the 25th conference on user modeling, adaptation and personalization. ACM, Bratislava, pp 408–412. https://doi.org/10.1145/3099023.3099112
29. Agus Santoso H et al (2018) Dinus intelligent assistance (DINA) chatbot for university admission services. In: 2018 international seminar on application for technology of information and communication. IEEE, Semarang, pp 417–423. https://doi.org/10.1109/ISEMANTIC.2018.8549797
30. Chun Ho C, Lee HL, Lo WK, Lui KFA (2018) Developing a chatbot for college student programme advisement. In: Wang FL, Iwasaki C, Konno T, Au O, Li C (eds) 2018 international symposium on educational technology (ISET). IEEE, Osaka, pp 52–56. https://doi.org/10.1109/ISET.2018.00021
31. Carayannopoulos S (2018) Using chatbots to aid transition. Int J Inf Learn Technol 35(2):118–129. https://doi.org/10.1108/IJILT-10-2017-0097
32. Singh J, Joesph MH, Jabbar KBA (2019) Rule-based chabot for student enquiries. J Phys Conf Ser 1228(1). https://doi.org/10.1088/1742-6596/1228/1/012060
33. Khin NN, Soe KM (2020) University chatbot using artificial intelligence markup language. In: 2020 IEEE conference on computer applications (ICCA). IEEE, Yangon, Myanmar. https://doi.org/10.1109/ICCA49400.2020.9022814
34. Patel NP, Parikh DR, Patel DA, Patel RR (2019) AI and web-based human-like interactive university chatbot (UNIBOT). In: 2019 3rd international conference on electronics, communication and aerospace technology (ICECA). IEEE, Coimbatore, pp 148–150. https://doi.org/10.1109/ICECA.2019.8822176
35. Verleger M, Pembridge J (2018) A pilot study integrating an AI-driven chatbot in an introductory programming course. In: 2018 IEEE frontiers in education conference (FIE), vol 2018. IEEE, San Jose, pp 1–4. https://doi.org/10.1109/FIE.2018.8659282
36. Mekni M, Baani Z, Sulieman D (2020) A smart virtual assistant for students. In: Petkov N, Strisciuglio N, Travieso-Gonzalez CM (eds) Proceedings of the 3rd international conference on applications of intelligent systems. ACM, Las Palmas de Gran Canaria, Spain. https://doi.org/10.1145/3378184.3378199
37. Vahdat M, Oneto L, Anguita D, Funk M, Rauterberg M (2015) Educational process mining (EPM): a learning analytics data set data set. https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set
38. Chung JY, Lee S (2019) Dropout early warning systems for high school students using machine learning. Child Youth Serv Rev 96:346–353. https://doi.org/10.1016/j.childyouth.2018.11.030
39. Limsathitwong K, Tiwatthanont K, Yatsungnoen T (2018) Dropout prediction system to reduce discontinue study rate of information technology students. In: 2018 5th international conference on business and industrial research (ICBIR). IEEE, Bangkok, pp 110–114. https://doi.org/10.1109/ICBIR.2018.8391176

# Multiprocessing Implementation for Building a DNA $q$-gram Index Hash Table


**Candace Claire Mercado, Aaron Russell Fajardo, Saira Kaye Manalili, Raphael Zapanta, and Roger Luis Uy**

**Abstract** Over the past few years, next-generation sequencing has become an invaluable technology for numerous applications in the field of genomics. The success of these applications are dependent on the performance of each phase in the genomic sequence pipeline, which starts with read mapping. However, read mapping is computationally intensive since it requires mapping billions of reads to numerous locations in a large reference genome. Building a $q$-gram index hash table has proven to be an efficient alternative to reduce the repetitive scanning of the reference during the verification step. A $q$-gram index hash table stores the locations of each $q$-gram in the reference genome. To accelerate the process of building this data structure and to exploit the multi-core architecture, instructions can be executed in parallel and distributed to multiple CPU cores. This paper performs a comparison analysis between the sequential and multiprocessing implementation of the index build time of the three methods for building a $q$-gram index hash table. The implementation results show that all multiprocessing versions are faster than sequential ones, with speedups ranging from 1.53 to 2.57. Although the open addressing method yields the fastest index build time, the best speedup is achieved by the minimizer-based method.

**Keywords** Bioinformatics · Read mapping · Index-based hash table

C. C. Mercado (✉) · A. R. Fajardo · S. K. Manalili
College of Computer Studies, De La Salle University, Manila, Philippines
e-mail: candace_mercado@dlsu.edu.ph

A. R. Fajardo
e-mail: aaron_fajardo@dlsu.edu.ph

S. K. Manalili
e-mail: saira_manalili@dlsu.edu.ph

R. Zapanta · R. L. Uy
Computer Technology Department, De La Salle University, Manila, Philippines
e-mail: raphael_zapanta@dlsu.edu.ph

R. L. Uy
e-mail: roger.uy@dlsu.edu.ph



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_16

## 1 Introduction

The advent of massively parallel sequencing or next-generation sequencing (NGS) platforms has proven to be an invaluable milestone in the field of genomics. These platforms paved the way for a multitude of new opportunities for researchers such as understanding the human genome diversity among populations [4], analyzing the genomic variants of ape species [17] and profiling of human diseases [15]. Despite the breakthroughs presented, these new sequencing platforms have shown to intensify the computational burden of genomic analysis [24]. NGS platforms generate a large amount of short DNA segments (also known as reads) which are later mapped to a known reference genome. This read mapping process is computationally expensive because billions of reads are being aligned to various locations of a large reference genome (i.e., human genome has 3.2 billion base-pairs). Another challenge is that each read may contain edits such as substitutions, insertions and deletions. This makes read mapping process an approximate string searching problem solvable using algorithms with quadratic-time complexity [3].

During the read mapping process, dynamic programming algorithms, such as Levenshtein edit distance [10] and Smith-Waterman [21], are typically employed in a verification step which measures the similarity between each read and candidate locations in a reference genome. However, these algorithms are inefficient especially given the amount of sequences that need to be processed. To significantly improve the read mapping process, most recent works utilize an approach called the filter-verification paradigm [1, 22, 25]. This approach uses a string-matching algorithm to compare the $q$-grams (subsequences with length $q$) of a read with those of a known reference genome. The dissimilar sequences are immediately filtered out and only the pool of candidate matches are later sent to the computationally intensive verification step [18].

As mentioned, the reference genome is often very large and traversing it repeatedly throughout the string matching process of the filtering step would be inefficient. Hence, a $q$-gram index hash table is generated. This data structure stores all occurrences of the $q$-grams present in the reference genome. Each $q$-gram serves as a basis for the hash key which subsequently determines where that $q$-gram will be stored in the hash table. Various works implementing a hash table have demonstrated to leverage open addressing schemes [16, 26, 27]. As for generating a $q$-gram index hash table, the latest work is RazerS3 [22], a read mapper based on counting $q$-grams which supports two filtering implementations (i.e., SWIFT and pigeonhole) and shared-memory parallelization. Each filter uses an open addressing $q$-gram index and the entire filter is being executed in parallel. The use of minimizers has also shown versatility in various applications such as $q$-gram counting [5, 7], metagenomics [13, 23] and read mapping. MashMap [8] and Minimap2 [12] are two read mappers which utilize a minimizer-based $q$-gram index hash table. Both collect minimizers from the reference genome and load these into a hash table. MashMap, however, builds a hierarchical index based on different window sizes. Another hash table implementation called direct addressing was also designed to build a $q$-gram index

[6]. This scheme, however, may not be optimal for larger values of $q$ as it stores all possible $q$-gram combinations. Given these works, there are currently none which employ multiprocessing techniques for generating direct addressing, open addressing and minimizer-based $q$-gram index hash tables.

Parallel implementations may be designed to further speedup the filtering step and fully take advantage of a multi-core system. Multiprocessing dwells on executing one task for every CPU core and multiple tasks executing in parallel on multiple CPU cores. An algorithm such as building the index hash table can be divided into subtasks and assign each to the available cores. This may present significant improvement in speed compared to sequential execution [9]. For this reason, in this paper, the goal is to perform a comparison analysis between the sequential and multiprocessing implementation for building a $q$-gram index hash table. The comparison will be in terms of index build time and will be conducted for three different index-building methods.

## 2 Index-Based Hash Table

The index-based hash table is a data structure built based on $q$-grams as the index and permits a more efficient retrieval of all $q$-gram occurrences from a known reference genome [14]. A $q$-gram is a sequence of $q$ characters from an alphabet $\Sigma$ where $\Sigma = \{A, C, G, T\}$ for DNA sequences. For example, the reference string ACGTT will yield the following overlapping 2-grams: AC, CG, GT and TT [20]. A $q$-gram is typically converted to its rank or its equivalent weighted value. This can be calculated by assigning A with a value of 0, C with a value of 1, G with a value of 2 and T with a value of 3. Thus, given a 2-gram CG, the corresponding rank value will be 6 (i.e., $1 * 4^1 + 2 * 4^0$). The following subsections discuss in detail three methods for building an index-based hash table: direct addressing, open addressing and minimizer-based.

### *2.1 Direct Addressing q-gram Index Hash Table*

The direct addressing $q$-gram index method [6] consists of two tables: the directory table and the position table. The directory table stores for each possible $q$-gram combination the starting location of that $q$-gram in the position table. A hashing function is not necessary since there is an entry for all possible $q$-grams (i.e., $4^q$) and each q-gram will yield a unique rank value. The position table stores the actual location where each $q$-gram occurs in the reference genome of length $n$. The number of entries in this table is equal to the number of $q$-grams present in the reference genome (i.e., $n - q + 1$). This method yields a time complexity of $O(4^q + n)$ since it populates both the directory and position tables. Figure 1 shows a detailed example of the direct addressing $q$-gram index hash table method given a reference string ACGTTAACTC and $q = 2$.

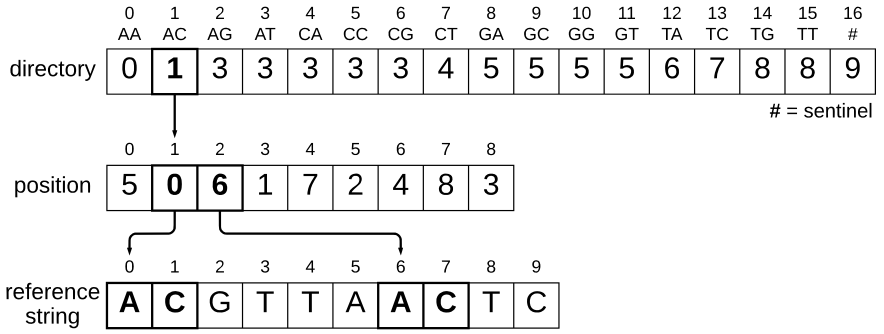



**Fig. 1** Direct addressing $q$-gram index hash table

## 2.2 Open Addressing q-gram Index Hash Table

Given the exponential space needed for the direct addressing method, memory usage quickly becomes a limiting factor for large $q$. However, not all $q$-gram combinations need to have an entry in the directory table since the maximum number of $q$-grams present in a reference genome is $n - q + 1$. The open addressing method [6] adds a new table called code table where each $q$-gram rank value is hashed into. This maps only the $q$-grams present in the reference genome to their respective positions in the directory table. It has a size of $\lfloor \alpha^{-1} n \rfloor$ where $\alpha$ is the fixed load factor with $0 < \alpha \leq 1$. A modulo operation may be used as a pseudo-random hash function and collisions may be handled with quadratic probing. Similar to direct addressing, the open addressing method generates a directory table and a position table exhibiting the same functionality. The size of the directory table, however, is reduced to the size of the code table plus one. This omits the exponential behavior of the memory consumption in a direct addressing hash table. As for the time complexity, this method yields an $O(2 + \alpha^{-1})n$ since it populates all three tables to build the index. Figure 2 shows a detailed example of the open addressing $q$-gram index hash table method given a reference string ACGTTAACTC, $q = 2$ and $\alpha = 0.8$.

## 2.3 Minimizer-Based q-gram Index Hash Table

The minimizer-based method [19], as the name suggests, was presented to minimize the storage requirements of $q$-gram index hash tables. This method stores the representative $q$-gram (i.e., minimizer) from each window of $w$ consecutive $q$-grams, where 'consecutive' means that each $q$-gram contained by that particular window is shifted to the right by one character. Given $w$ consecutive $q$-grams, the length of each window will be $w + q - 1$. Choosing the representative $q$-gram can be done by obtaining the smallest rank value or any other ordering depending on the require-

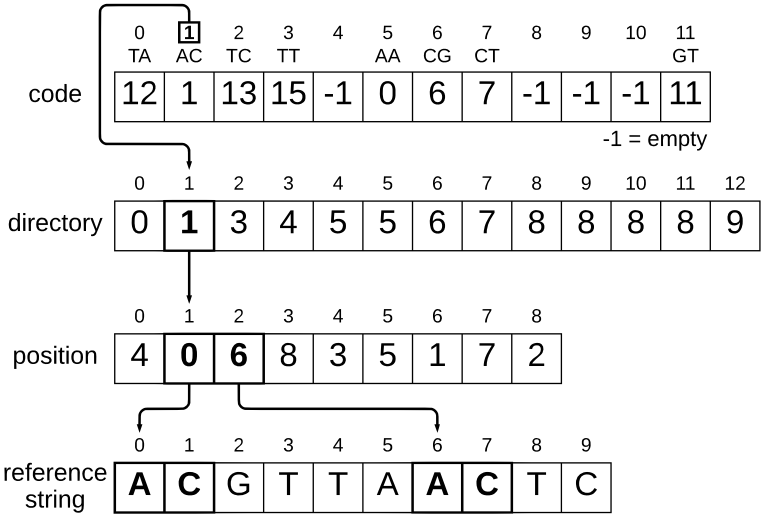



**Fig. 2** Open addressing $q$-gram index hash table

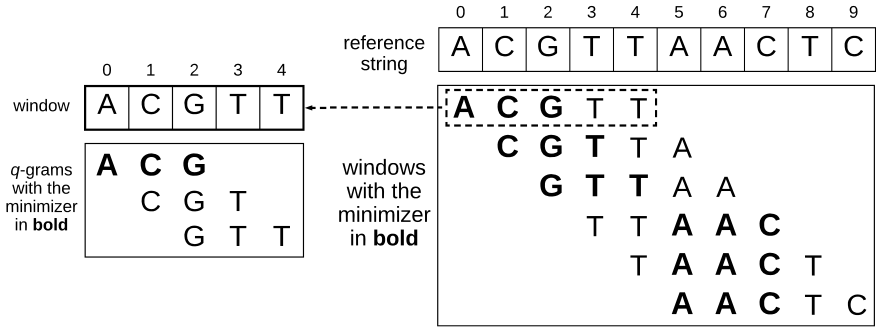



**Fig. 3** Selecting the minimizers within widows in a reference string

ment. As shown in Fig. 3, given $q = 3$ and $w = 3$, the smallest ranking $q$-gram is ACG and therefore it becomes the minimizer for the window ACGTT.

The method of selecting a minimizer within a window is performed to all windows present in a reference genome, as illustrated in Fig. 3. The figure shows three consecutive windows choosing the same minimizer AAC. This behavior is fundamental to how minimizers significantly reduce the storage requirement for a $q$-gram index hash table. Instead of storing all 8 $q$-grams present in the reference string, only 4 will be stored. Once a minimizer is selected from a window, its rank value goes through an invertible integer hash function. The hashed value acts as the key in the hash table containing the positions of the windows where this minimizer was selected.

To build the index, the minimizer-based method performs two loops: an outer loop which splits the reference genome into windows and an inner loop which selects the minimizer from each window. This translates to a time complexity of $O(n - 2q)q$.

## 3 Methodology

Both sequential and multiprocessing versions for the three index-building methods are implemented in C++. The multiprocessing version is achieved by utilizing OpenMP,[1] a library which supports shared-memory parallelism for C++ programs. It allows the conversion of programs from sequential to multiprocessing with minimal code modifications by using simple compiler directives. OpenMP's directives tell the compiler which instructions must be executed in parallel and how to distribute these instructions among the available threads [2]. Although this simplifies the task at hand, code analysis still has to be performed by the programmer to identify if code blocks can be executed in parallel. Algorithm 1 shows the multiprocessing implementation of the minimizer-based method through the use of OpenMP directives. The construct $\#pragma\ omp\ parallel\ for$ (at line 4) marks the start of a parallel region containing a for loop. This is convenient if there are no dependencies among instructions within the loop body. Since the iterations of the outer loop involve selecting a minimizer within each window, each iteration can be executed independently and distributed among the available threads. Furthermore, the $\#pragma\ omp\ critical$ construct (at line 18) specifies that the enclosed code block is a critical section. In other words, it must execute an instruction one thread at a time. This is used to safely allow multiple threads to access and update the array of minimizers.

**Algorithm 1** OpenMP implementation of the minimizer-based method.

```
1: ref ← reference string
2: n ← length of ref
3: wLength ← w + q − 1
4: loop#pragma omp parallel for
5:     for i ← 0 to n − wLength do
6:         window ← ref from i to i + wLength − 1
7:         for j ← 0 to wLength − q do
8:             qGram ← window from j to j + q − 1
9:             rankVal ← extractRank(qGram)
10:            hashVal ← intHash(rankVal)
11:            if hashVal < min then
12:                min ← hashVal
13:                minzer ← qGram
14:            end if
15:        end for
16:        rankValMin ← extractRank(minzer)
17:        hashValMin ← intHash(rankValMin)
18:        loop#pragma omp critical
19:            push i to minimizers[hashValMin]
20:        end loop#pragma
21:    end for
22: end loop#pragma
```

For the multiprocessing implementation of the direct addressing method, as shown in Algorithm 2, there are three loops involved. The first loop iterates, in parallel, through each $q$-gram and converts each to its rank value. The succeeding instruction (at line 8), however, increments the directory table at the position corresponding to

[1] https://www.openmp.org/.

the rank value. Since there can be multiple $q$-gram instances, multiple iterations can update a single memory location. Thus, the instruction is enclosed in #*pragma omp atomic* (at line 7). Although the *atomic* and *critical* constructs both prevent race conditions, *atomic* exhibits a substantially lower overhead and is limited to only a restricted set of operations.

**Algorithm 2** OpenMP implementation of the direct addressing method.

```
1: ref ← reference string
2: n ← length of ref
3: loop#pragma omp parallel for
4:     for i ← 0 to n − q do
5:         qGram ← ref from i to i + q − 1
6:         rankVal ← extractRank(qGram)
7:         loop#pragma omp atomic
8:             increment dirTable[rankVal] by 1
9:         end loop#pragma
10:    end for
11: end loop#pragma
12: for i ← 1 to size of dirTable − 1 do
13:     dirTable[i] ← dirTable[i] + dirTable[i − 1]
14: end for
15: loop#pragma omp parallel for
16:     for i ← n − q to 0 do
17:         qGram ← ref from i to i + q − 1
18:         rankVal ← extractRank(qGram)
19:         posIndex ← 0
20:         loop#pragma omp critical
21:             decrement dirTable[rankVal] by 1
22:             posIndex ← dirTable[rankVal]
23:         end loop#pragma
24:         posTable[posIndex] ← i
25:     end for
26: end loop#pragma
```

As for the second loop, it contains a single statement which cascades values throughout the directory table. It possesses a loop-carried dependency and therefore, is not implemented in parallel. The last loop employs the *critical* construct since there are two consecutive statements that need to be executed one thread at a time. On different iterations, the first instruction can decrement the element from a single array location while the second "snapshots" this decremented value and stores it in $posIndex$, which is later used as the key pointing to the location in the position table where a $q$-gram is stored.

The open addressing method involves three loops to build the hash table. The first loop builds the code table which stores the rank values of all $q$-gram occurrences in the reference genome. To do this, every iteration contains an if-else statement which checks if a location in the code table is empty. Only then can the rank value be stored. However, the multiprocessing structure of the loop can be vulnerable to data races (i.e., two threads accessing a shared memory location at the same time). This can be due to simultaneous checking of the code table from different points in the code segment. As a result, two iterations may be hashing into one code table position, disregarding the earlier assignment.

**Algorithm 3** OpenMP implementation of the open addressing method.

```
1: ref ← reference string
2: n ← length of ref
3: loop#pragma omp parallel for
4:     for i ← 0 to n − q do
5:         qGram ← ref from i to i + q − 1
6:         rank ← extractRank(qGram)
7:         hashVal ← rank mod codeTable size
8:         omp_set_lock()
9:         if codeTable[hashVal] ≠ −1 then
10:            j ← hashVal
11:            k ← 1
12:            while codeTable[j] ≠ (−1 and rank) do
13:                j ← (j + (k ∗ k)) mod codeTable size
14:                k ← k + 1
15:            end while
16:            codeTable[j] ← rank
17:            increment dirTable[j] by 1
18:            omp_unset_lock()
19:            codeTableIndices[i] ← j
20:        else
21:            codeTable[hashVal] ← rank
22:            increment dirTable[hashVal] by 1
23:            omp_unset_lock()
24:            codeTableIndices[i] ← hashVal
25:        end if
26:    end for
27: end loop#pragma
```

OpenMP's support functions for the mutex lock (at line 8, 18 and 23) are used to ensure that the code table locations are being accessed and modified one thread at a time, as shown in Algorithm 3. While locks may be similar to using *pragma* and *atomic* constructs, locks are about the data, not the code segment. For instance, given the *atomic* implementation (at line 7 in Algorithm 2), although it involves updating a shared element in the directory table as well, the only requirement is to prevent two processes from executing an update instruction simultaneously, regardless of what is being updated. The use *atomic* is a sufficient to yield correct results. In this method, however, a lock is deemed to be the best technique to prevent two processes from executing an update instruction simultaneously on a single memory location. This is to guarantee a "lock" on the element in the code table which is crucial to the correctness of the other loop iterations. The second and third loops have a multiprocessing implementation similar to direct addressing with the small difference in the third loop. The third loop iterates through *codeTableIndices* instead of each $q$-gram which means that rank extraction is omitted and code table indices are used to index the directory table.

The testing platform used is an Ubuntu 18.04 system with a quad-core Intel® Xeon® 2.60 GHz CPU (2 threads each core) and 16 GiB of memory. All implementations are run on a simulated and a real reference genome data set. The simulated data set has a length of 1,048,576 and contains randomly generated DNA characters. As for the real data set, the Human genome (GRCh38) chromosome 4 is used, which is obtained from the National Center for Biotechnology Information (NCBI) website.[2] It has a length of 189,752,667 and has all non-ACGT characters removed. Except for the direct addressing method, the index build time is documented for each $q$ value of 8, 10, 12, 14, 16, 18 and 20. Given memory limitations, direct addressing can only execute up to a $q$-gram size of 14. For the minimizer-based method, the $w$ value is set equal to the value of $q$. This allows every character in the reference string to be covered by at least one minimizer [19]. Moreover, this method passes each $q$-gram

[2] https://www.ncbi.nlm.nih.gov/genome/guide/human/.

rank value to an invertible integer hash function[3] first before selecting the minimizer within each window. The hash function hinders non-informative poly-A (i.e., AAAA), the smallest ranking sequence, to always be selected as minimizers [11].

## 4 Results and Discussion

The experimental results for the two data sets, summarized in Tables 1 and 2, show that all OpenMP-based implementations outperform sequential implementations for each of the three methods. The direct addressing method demonstrates a speedup from 2.06 for $q = 8$ down to 1.59 for $q = 14$ in the simulated dataset and 1.92 for $q = 8$ down to 1.68 $q = 14$ in the real dataset. As discussed in the previous section, the second loop iterating through the directory table contains dependencies and therefore, is implemented sequentially. As such, the speedup decreases as the directory table's size increases exponentially. For the open addressing method, the speedups range from 1.53 to 1.65 for the synthetic data set and 1.69 to 1.86 for the real data set. Despite having the lowest index build time, the speedup is lower in comparison to the other two methods (from $q = 8$ to $q = 12$). This is attributed to the use of OpenMP's lock functions especially since the process of resolving collisions is enclosed in it. The minimizer-based multiprocessing implementation resulted in the best speedup being at least 1.84 faster for the simulated data set and at least 2.51 times faster for the real data set. This due to a more optimized use of critical regions, as shown at line 18 in Algorithm 1 where there is only one instruction contained by the $critical$ construct inside the whole loop. For validation purpose, another simulated data set with a length of 4,194,304 (i.e., 4 MB) and another real data set *Escherichia coli* with a length of 5,498,578 (i.e., *E. coli*) was run (data not tabulated). The speedup results obtained are consistent. For direct addressing, speedups also decline from 1.78 ($q = 8$) to 1.63 ($q = 14$) for 4 MB and from 1.92 ($q = 8$) to 1.80 ($q = 14$) for *E. coli*. The speedups for open addressing range from 1.75 to 1.86 for 4MB and 1.76 to 1.84 for *E. coli*. The minimizer-based method still has the best speedups: ranging from 1.90 to 2.04 for 4MB and 1.92 to 2.02 for *E. coli*.

The speedups of the current implementations are affected by several types of overheads incurred due to the utilization of OpenMP. There are sequential overheads, which are essential to some computations to ensure that correct results are obtained. Moreover, there exist synchronization overheads resulting from the use of constructs such as $critical$ and $atomic$ as well as the support functions for locks. From this, load imbalance among synchronization points is a likely occurrence. In other words, certain threads become idle once they finish executing instructions and have to wait for the slower threads. Lastly, there are also parallelization overheads, which are incurred due to the use of OpenMP directives to designate parallel regions.

[3] https://naml.us/post/inverse-of-a-hash-function/.

**Table 1** Index hash table build time (in seconds) of the three methods in for the simulated data set ($n = 1,048,576$)

| | Direct addressing | | | Open addressing | | | Minimizer-based | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sequential | Multiprocessing | | Sequential | Multiprocessing | | Sequential | Multiprocessing | |
| $q$ | | | Speedup | | | Speedup | | | Speedup |
| 8 | 5.82 | 2.83 | 2.06 | 2.76 | 1.78 | 1.55 | 23.24 | 11.71 | 1.98 |
| 10 | 7.20 | 3.52 | 2.05 | 3.43 | 2.16 | 1.58 | 34.60 | 17.65 | 1.96 |
| 12 | 9.06 | 4.47 | 2.03 | 4.03 | 2.63 | 1.53 | 48.42 | 24.96 | 1.94 |
| 14 | 20.13 | 12.69 | 1.59 | 4.81 | 2.95 | 1.63 | 65.44 | 33.36 | 1.96 |
| 16 | N/A | N/A | N/A | 5.37 | 3.31 | 1.62 | 85.05 | 45.52 | 1.87 |
| 18 | N/A | N/A | N/A | 5.93 | 3.66 | 1.62 | 106.13 | 57.68 | 1.84 |
| 20 | N/A | N/A | N/A | 6.58 | 3.99 | 1.65 | 128.60 | 68.95 | 1.87 |

**Table 2** Index hash table build time (in seconds) of the three methods in for the real data set ($n = 189,752,667$)

| | Direct addressing | | | Open addressing | | | Minimizer-based | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sequential | Multiprocessing | | Sequential | Multiprocessing | | Sequential | Multiprocessing | |
| $q$ | | | Speedup | | | Speedup | | | Speedup |
| 8 | 900.63 | 468.40 | 1.92 | 491.44 | 289.96 | 1.69 | 5390.36 | 2144.25 | 2.51 |
| 10 | 1158.91 | 630.63 | 1.84 | 609.68 | 343.35 | 1.78 | 8247.21 | 3252.91 | 2.54 |
| 12 | 1406.76 | 775.34 | 1.81 | 749.07 | 419.92 | 1.78 | 11,564.16 | 4566.83 | 2.53 |
| 14 | 1629.84 | 972.53 | 1.68 | 887.98 | 484.82 | 1.83 | 15,478.96 | 6136.03 | 2.52 |
| 16 | N/A | N/A | N/A | 986.86 | 536.15 | 1.84 | 20,448.72 | 7968.93 | 2.57 |
| 18 | N/A | N/A | N/A | 1095.43 | 614.03 | 1.78 | 25,574.94 | 9939.75 | 2.57 |
| 20 | N/A | N/A | N/A | 1208.98 | 649.46 | 1.86 | 31,191.65 | 12,120.68 | 2.57 |

The index build time of all multiprocessing implementations for varying $q$-gram sizes are illustrated in Figs. 4 and 5. In terms of selecting the fastest method for building a $q$-gram index hash table, open addressing outperforms the other two methods. While its index build time is independent of the $q$-gram size, resolving collisions results in incremental increase in time as the $q$ value increases. Direct addressing, as previously discussed, grows exponentially as $q$ increases. This behavior is distinguishable for the simulated data set but not as much for the real data set due to limitations in executing $q$ values over 14. As for the minimizer-based method, its index build time is dependent on the value of $q$ and performs the worst for $q = 8$ to $q = 14$. For larger values of $q$, however, it will be surpassed by direct addressing method's exponential build time.

Figure 6 shows the $q$-gram index hash tables generated after executing the three multiprocessing implementations on a smaller input. The 2-gram AG occurs at positions 8, 10 and 18 in the reference sequence TATGCACCAGAGTATGGAAG, which

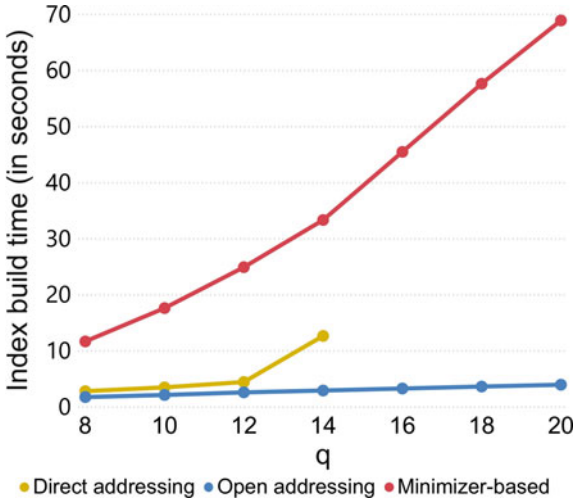



**Fig. 4** Multiprocessing implementation index build time vis-a-vis $q$-gram size for the simulated data set ($n = 1,048,576$)

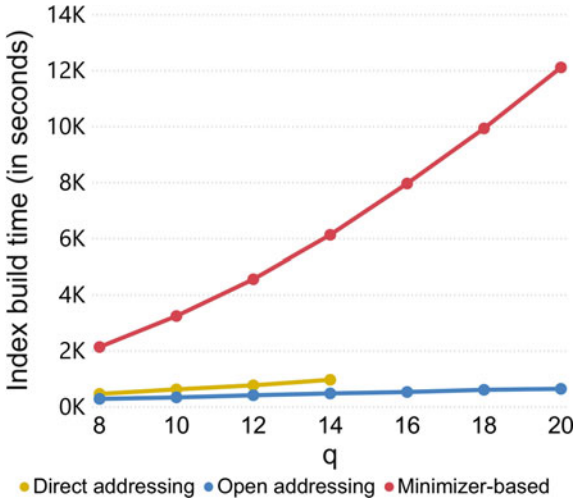



**Fig. 5** Multiprocessing implementation index build time vis-a-vis $q$-gram size for the real data set ($n = 189,752,667$)

has a length of 20. For the minimizer-based method, AG's rank value 2 is passed into the invertible hash function which yields a value of 9. The figure shows that at position 9 in the index, the windows which select AG as their minimizer are found at positions 8, 9, 7 and 18.

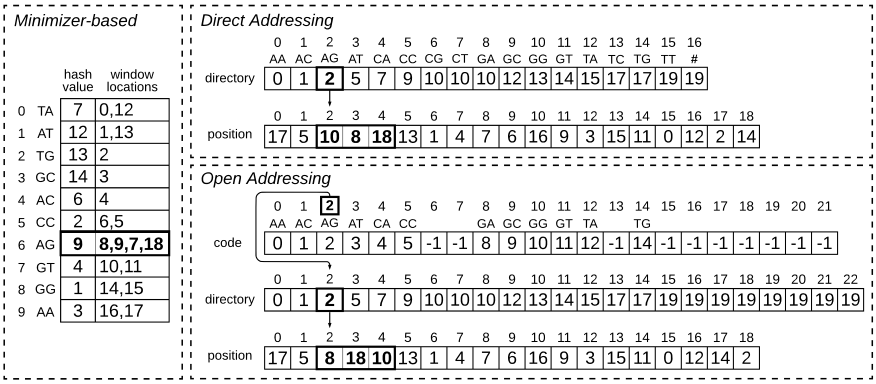



**Fig. 6** Resulting $q$-gram index hash tables given $n = 20$, $q = 2$, and $\alpha = 0.9$

## 5 Conclusion and Future Work

In this paper, an OpenMP-based multiprocessing implementation was employed for three methods of building a $q$-gram index hash table. The implementation results were presented as a comparison of the index build time between the sequential and the multiprocessing versions. Based on the outcomes of this study, the minimizer-based method achieves the highest speedup. Its implementation makes it least 2.51 faster for the real data set. This is due to its implementation's reduced use of synchronization techniques. Speedups are not as significant for the other two methods due to loop-carried dependencies and OpenMP's synchronization overheads. Overall, exploiting a multi-core system through an OpenMP-based parallel implementation yields a better performance for building an index hash table compared to a sequential one. These speedups, however, are only run on a single quad-core processor system. Therefore, future work will involve conducting a performance study of executions on varying number of cores. Other libraries for parallelism as well as distributed computing can also be explored. Additionally, other multiprocessing techniques can be investigated to identify a better implementation for each index-building method.

## References

1. Alser M, Hassan H, Kumar A, Mutlu O, Alkan C (2019) Shouji: a fast and efficient pre-alignment filter for sequence alignment. Bioinformatics 35(21):4255–4263
2. Barlas G (2014) Multicore and GPU programming: an integrated approach. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
3. Canzar S, Salzberg SL (2017) Short read mapping: an algorithmic tour. Proc IEEE Inst Electr Electron Eng 105(3):436–458
4. Consortium GP (2012) An integrated map of genetic variation from 1092 human genomes. Nature 491(7422):56–65

5. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A (2015) KMC 2: fast and resource-frugal k-mer counting. Bioinformatics 31(10):1569–1576
6. Elloumi M (ed) Algorithms for next-generation sequencing data: techniques, approaches, and applications. Springer (2017). https://doi.org/10.1007/978-3-319-59826-0
7. Erbert M, Rechner S, Müller-Hannemann M (2017) Gerbil: a fast and memory-efficient k-mer counter with gpu-support. Algor Mol Biol 12: https://doi.org/10.1186/s13015-017-0097-9
8. Jain C, Dilthey A, Koren S, Aluru S, Phillippy AM (2018) A fast approximate algorithm for mapping long reads to large reference databases. J Comput Biol 25(7):766–779
9. Langenkämper D, Jakobi T, Feld D, Jelonek L, Goesmann A, Nattkemper TW (2016) Comparison of acceleration techniques for selected low-level bioinformatics operations. Front Genet 7:5. https://doi.org/10.3389/fgene.2016.00005
10. Levenshtein V (1966) Binary codes capable of correcting deletions. Insertions and Reversals. Soviet Physics Doklady 10:707
11. Li H (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32(14):2103–2110
12. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191
13. Li K, Lu Y, Deng L, Wang L, Shi L, Wang Z (2020) Deconvolute individual genomes from metagenome sequences through short read clustering. PeerJ 8:e8966
14. Liu J, Chen Q, Zhang C (2015) K-mer index of DNA sequence based on hash algorithm. Int J Comput Sci Appl 5(4):19–28
15. Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10(21):2319–2328
16. Nielsen JP, Karlsson S (2016) A scalable lock-free hash table with open addressing. SIGPLAN Not. 51(8). https://doi.org/10.1145/3016078.2851196
17. Prado-Martinez J et al (2013) Great ape genetic diversity and population history. Nature 499(7459):471–475
18. Reinert K, Langmead B, Weese D, Evers DJ (2015) Alignment of next-generation sequencing reads. Ann Rev Genom Hum Genet 16(1):133–151. https://doi.org/10.1146/annurev-genom-090413-025358
19. Roberts M, Hayes W, Hunt BR, Mount SM, Yorke JA (2004) Reducing storage requirements for biological sequence comparison. Bioinformatics 20(18):3363–3369
20. Salmela L, Tarhio J, Kytöjoki J (2007) Multipattern string matching with q-grams. ACM J Exp Algorithm 11:1.1-es (2007). https://doi.org/10.1145/1187436.1187438
21. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
22. Weese D, Holtgrewe M, Reinert K (2012) RazerS 3: faster, fully sensitive read mapping. Bioinformatics 28(20):2592–2599. https://doi.org/10.1093/bioinformatics/bts505
23. Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. bioRxiv (2019). https://doi.org/10.1101/762302, https://www.biorxiv.org/content/early/2019/09/07/762302
24. Xin H, Greth J, Emmons J, Pekhimenko G, Kingsford C, Alkan C, Mutlu O (2015) Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping. Bioinformatics 31(10):1553–1560
25. Xin H, Lee D, Hormozdiari F, Yedkar S, Mutlu O, Alkan C (2013) Accelerating read mapping with FastHASH. BMC Genom 14(S1):S13. https://doi.org/10.1186/1471-2164-14-S1-S13
26. Yaniv I, Tsafrir D (2016) Hash, don't cache (the page table). SIGMETRICS Perform. Eval Rev 44(1):337–350. https://doi.org/10.1145/2964791.2901456
27. Zheng T, Zhang Z, Cheng X (2020) Saha: a string adaptive hash table for analytical databases. Appl Sci 10(6):1915. https://doi.org/10.3390/app10061915

# Predicting Chart Difficulty in Rhythm Games Through Classification Using Chart Pattern Derived Attributes


**Arturo P. Caronongan III and Nelson A. Marcos**

**Abstract** Rhythm games are music-themed games that challenge players' sense of rhythm and reaction skills. One such popular rhythm-based video game is Dance Dance Revolution, where players perform steps on a dance platform that is synchronized with music as directed by on-screen step charts. An issue that exists, not just in Dance Dance Revolution, but in rhythm games in general is the estimation of a chart's difficulty level. While many methods and studies exist in generating and predicting chart attributes, there is no clear methodology existing in determining the optimal difficulty of a given chart. This paper aims to address the aforementioned issue in the game of Dance Dance Revolution by proposing a methodology that involves extracting patterns and common attributes in step charts that enable more accuracy in determining a chart's difficulty level. The resulting methodology achieved an average True-Positive rating of 0.683 and an overall model accuracy of 74.82% for classifying charts according to levels in Dance Dance Revolution.

**Keywords** Machine learning · Classification · Data mining · Entertainment computing

## 1 Introduction

### *1.1 Overview of the Current State of Technology*

Dance Dance Revolution (DDR) is a rhythm-based video game with millions of players worldwide [1]. In the game, players perform steps atop a dance pad by following prompts from an on-screen stepchart (illustrated in Fig. 1) that are synced with music that plays. The dance pad contains up, down, left, and right arrows which can be in one of the four states: on, off, hold, or release. Since the four arrows can

A. P. Caronongan III (✉) · N. A. Marcos
De La Salle University Manila, Manila, Philippines
e-mail: arturo.caronongan@delasalle.ph

N. A. Marcos
e-mail: nelson.marcos@dlsu.edu.ph



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_17

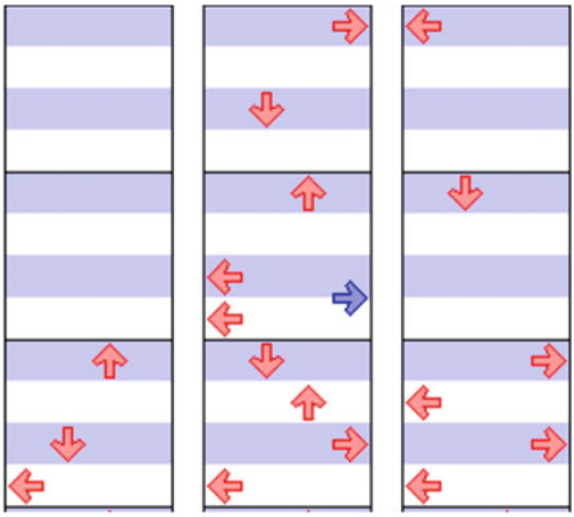


**Fig. 1** An excerpt of a stepchart of a song. Stepcharts are a series of arrows that the user will have to step to in a given rhythm. A stepchart's difficulty increases depending on the stamina technique required to successfully execute given patterns

be activated or released independently, there are 256 possible step combinations at any instant [2, 3].

A player's score depends on hitting the correct buttons at the correct time and rhythm. Stepcharts vary in difficulty with harder charts containing more steps and more complex sequences that normally require more movement [3]. It is due to this that stepcharts exhibit rich structures and complex semantics to ensure that step sequences are both challenging and enjoyable to players of varying levels. There are varied types of players, namely beginners to competitive players, for the rhythm-based video games and such is true for Dance Dance Revolution. As a result, multiple charts for each difficulty level are prepared for the same song. The game creators manually compose these multiple charts in general.

The difficulty level of the charts is discretely composed, thus there is sometimes too much distance between the difficulty levels. Some of the players, especially average players, are not satisfied enough with the charts because the charts in the appropriate difficulty level for the player are not prepared. A study named Dance Dance Gradation [4] aimed to solve this issue that was not addressed in the aforementioned study. In the study, the system learned the relationships between difficult and easy charts based on the deep neural network using a dataset of dance charts with different difficulty levels as the training data. The difficulty chart automatically would be adapted to easier charts through the learned model. As mixing multiple difficulty levels for the training data, the generated charts should have each characteristic of difficulty level. The user can obtain the charts with intermediate difficulty level between two different levels.

### *1.2 Problem and Limitations of Existing Studies*

Despite the approaches undertaken to add quantifiable values for step charts and studies conducted for generating step charts, there exists an issue of accurately making an estimate as to what a chart's difficulty level is. Although there are metrics that have been used to calculate different values of a chart, the final rating is still relatively decided by the main designers of the game itself. Even after the final ratings have been decided, there has been a history of charts being re-rated due to the average score being deemed lower or higher than the average score for charts of that particular level or the existence of new charts.

This paper aims to discuss a methodology that will be able to provide a proposed solution to the existing limitation.

### *1.3 Scope and Limitations of the Study*

For this research, the game of Dance Dance Revolution (DDR) will be used as the reference rhythm game due to existing literature being provided with regards to studies that have been conducted with regards to chart generation and in terms of health benefits. Likewise, it is for this game where datasets representing the charts of the game are available through open source communities as of this writing. As players normally report discrepancies between ratings of charts between charts that are rated Level 13 to Level 19, therefore it is charts that contain these difficulty levels that will be studied in this paper. Although the charts are rated from 1 to 20 in the game of DDR, with charts rated as Level 20 being the most difficult, charts rated Level 1 to 12 aren't normally debated upon due to the absence of technical measurements not observed in charts rated with a difficulty of Level 13 to 19. Likewise, most of the charts that get re-rated are normally within the range between Levels 13 to 19 [5].

DDR's ratings are in the form of a classification, thus classification algorithms will be used for this study, but logistic (not linear) regression will also be considered. Likewise, each different level will be treated as a nominal label. More of this will be explained in the succeeding section. Finally, throughout this paper, the word "chart" will be used to refer to a stepchart, the definition of which was explained in Sect. 1.1.

## 2 Review of Related Literature

### *2.1 Chart Structure and Step Patterns*

In rhythm games the player hits notes by pressing buttons on an input device. The goal is to hit notes in time to the music, which can be represented as a sequence S of note, time pairs:

$$S := \{s_i = (n_i, t_i), n_i \in C, t_i \in \mathbb{R}, t_1 < t_2 \ldots t_N\}_{i=1}^{N}, \quad (1)$$

where C is a finite set of possible notes, and N the total number of notes in the song. Chords, or sets of in-game notes that occur at the same ti, are represented as a distinct note [6].

Stepcharts in DDR are normally parsed through an open source Dance Dance Revolution emulator named Stepmania through. SM files [7, 8]. SM files are text files that contain information about the song's contents, such as the title of the song, the BPM, the audio file, and any other related song attributes related to the game. Most importantly, SM files are used to represent a song's corresponding chart through a collection of texts, as shown in Fig. 2.

```
0000 // 1st beat of the measure
0000 // 2nd beat of the measure
0000 // 3rd beat of the measure
…
0000 // n'th beat of the measure
,    // comma representing the end of the measure
```

A series of patterns that may exist in charts have been labeled with terms. These patterns are coined as the following with Fig. 3 illustrating the aforementioned patterns. *Steps* (1) are a solitary step, where a single arrow in any direction is a step. Multiple steps are placed together in a consecutive order, forming certain patterns. *Jumps* (2) are simultaneous steps that come in six different variations in DDR, which are done with two feet hitting both steps at the same time. *Freeze* (3) is performed by stepping and holding down on a note until the freeze portion end. *Staircase* (4) is a


**Fig. 2** Representation of a DDR chart's measure. Each measure contains n lines, where n = number of beats in a measure, and 0 s indicate no step while a 1 indicates a normal step. 2, on the other hand indicates a freeze arrow while 3 indicates the release of a previous freeze arrow

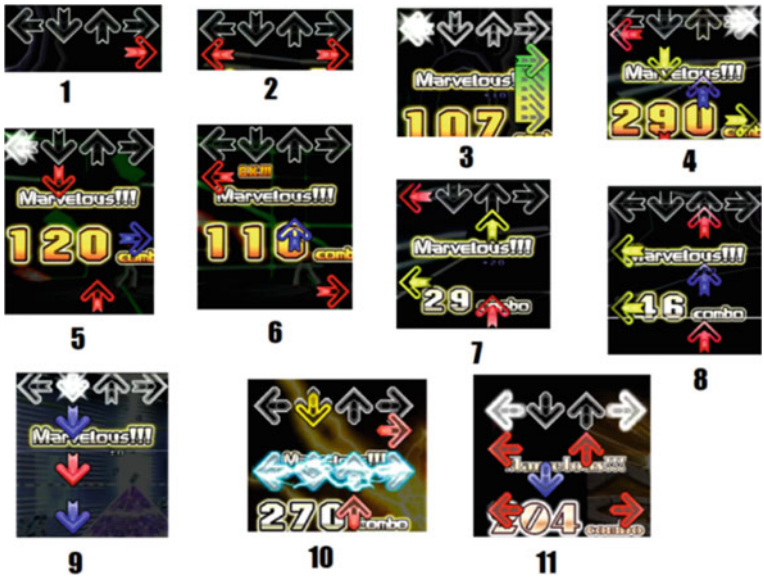



**Fig. 3** Existing patterns that encompass a song's step chart. It is a consensus among players that the presence of these patterns may influence a chart's difficulty due to the advanced technique required to execute some of them at high speeds [7]

pattern that occurs when players hit all arrows once going from left to right or right to left. *Candle* (5) patterns take more movement than other patterns to execute because of how they make a player's foot on the first step go to the opposite panel on the third step. Some examples of candle patterns are Up Left Down and Down Right Up. *Crossovers* (6) are complex patterns which cause either a player's left foot to step on the right arrow, or the player's right foot to step on the left arrow. Example patterns are: Left Down Right, Left Up Right; and the reverse. *Gallops* (7) are two different arrows in quick succession, typically 1/16th note apart. *Drills* (8) are at least five consecutive notes, typically 1/16th notes (or 1/8th notes at high speed) alternating on the same two arrows. *Jacks* (9) are a series of two or more consecutive notes on the same arrow. *Shocks* (9) are arrows that occupy all panels simultaneously which must be avoided by jumping or by placing both feet in the middle panel. Finally, *Step Jumps* (10) are Patterns that involve a single step that precedes or occurs after a Jump. Normally, these are classified as 8th singular steps followed by a jump [7].

## 2.2 Groove Radar and Groove Radar Values

The game has an official algorithm for calculating specific attributes broken down into the following: Stream, Voltage, Air, Freeze, and Chaos. The Groove Radar, as depicted in Fig. 3, is a graphical representation of a song's step contents using a

pentagon shape. Originally meant to be the successor to the traditional Feet/Number Level system, both the Feet/Number Level and Groove Radar have been used since then. While not meant to be an accurate representation of a stepchart's difficulty, it provides players with a graphical notation of the stepchart's contents [9].

Stream is denoted as the amount of sets of steps that are right after one another (step density) in the song. It is actually calculated as the average number of steps per minute. Voltage refers to the peak density of the steps of the song. It is the highest density of arrows that appear on the screen at once in one measure (four beats). Air refers to the total number of jumps and Shock Arrows in the chart. Freeze is denoted as the total length of the chart's Freeze Arrows. Chaos refers to the amount of steps that do not match well with the beat of the song [9].

## 3 Preliminary Data Analysis

### *3.1 Description of the Dataset*

The dataset consists of 564 entries describing the Groove Radar Values generated from the equations mentioned in [9]. The dataset is unbalanced, with a total of 196 level 13 charts, 108 level 14 charts, 100 level 15 charts, 74 level 16 charts, 41 level 17 charts, 39 level 18 charts, and only 6 level 19 charts. The reason for the small amount in the dataset is due to the entries being limited to only official charts that are available within the game. Likewise, performing data balancing techniques such as Random Oversampling, Random Undersampling, or Synthetic Minority Oversampling Techniques might produce values that represent charts that have not yet been developed, so this experiment decided to make use of actual data that already contains concrete representations of existing charts.

### *3.2 Correlation of Groove Radar Attributes to the Level*

Given the default analysis and attributes given, it can be clearly observed that Stream, Voltage, and Air are positively correlated to the level as indicated in Fig. 4. This is further evidenced in Table 1, which shows the difference between the average values of levels compared to the average value of charts one level below it. It can be observed that all were positively correlated with the exception of Freeze and Chaos (Fig. 5).

Further analysis observed that judging each chart by the Groove Radar values alone showed that there may be charts that are rated lower but have values higher than the average value of the next corresponding level. Table 2 displays the difference obtained from subtracting the MAX (M) value of a corresponding level from the Average (A) level of a level higher than the corresponding level.


**Fig. 4** The Groove Radar, which is used to visualize a chart's stream, chaos, freeze, air, and voltage via a pentagonal figure as taken from the Stepmania [9] interface. Ideally, higher level charts have a bigger Groove Radar due to the higher density of notes

**Table 1** Listing of the differences of the average Groove Radar values for difficulties one level apart

| Difference | Stream | Voltage | Air |
|---|---|---|---|
| Lv 13–14 | 9.71 | 15.88 | 2.30 |
| Lv 14–15 | 12.41 | 9.99 | 5.98 |
| Lv 15–16 | 14.33 | 9.50 | 4.30 |
| Lv 16–17 | 16.00 | 13.40 | 7.27 |
| Lv 17–18 | 17.22 | 12.15 | 10.35 |
| Lv 18–19 | 30.68 | 15.53 | 5.252 |

**Fig. 5** Correlation of the average Groove Radar values in determining a chart's difficulty and how they are positively correlated

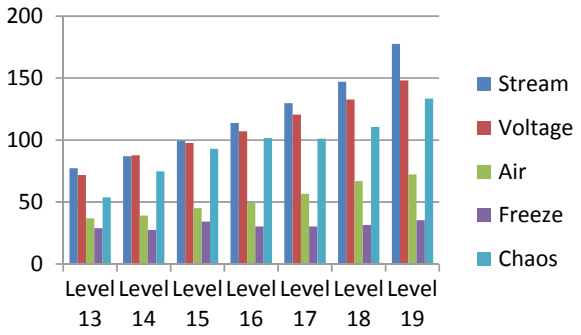


**Table 2** Listing of the differences of the average Groove Radar value for one level lower

| Mean–Ave | Stream | Voltage | Air | Freeze | Chaos |
|---|---|---|---|---|---|
| M(13)–A(14) | 10.037 | 18.38 | 160.94 | 92.47 | 44.27 |
| M(14)–A(15) | 22.63 | 79.39 | 66.96 | 59.76 | 71.07 |
| M(15)–A(16) | 29.30 | 51.89 | 78.66 | 70.70 | 17.41 |
| M(16)–A(17) | 29.29 | 20.49 | 51.39 | 70.78 | 98.85 |
| M(17)–(18) | 15.08 | 44.33 | 48.05 | 49.46 | 61.38 |
| M(18)–A(19) | 22.40 | 111.80 | 45.80 | 64.60 | 41.60 |

Under ideal scenarios, it is best if the values of all relevant attributes (deemed to be Stream, Voltage, Air) of a particular level will always be higher than the values of charts of a lower level. However, that is not the case. Here, the table displays positive values whenever the AVERAGE value of a particular level is subtracted from the MAX value of one level below it.

### *3.3 Classification Using Groove Radar Values*

While determining the difficulty ratings of a chart has many factors, it is eventually the game designer's decision on what a particular chart's level is. However, the Groove Radar Values may give an estimate on what a chart's contents are. To check whether these values are appropriate in determining a chart's given difficulty value, preliminary classification experiments were performed on the data set. Classification models with a tenfold cross validation used were the Multilayer Perceptron (MLP), Logistic Regression (LR), Naïve Bayes (NB), and the K Nearest Neighbor (Knn) using WEKA [10]. The different models and their performances are indicated in Table 3, describing the accuracy (ACC), Cohen's Kappa statistic (KP), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Since the dataset might be a bit too unbalanced to rely on accuracy alone, Table 4 provides an in depth preview of the corresponding True-Positive, False-Positive, Precision, Recall, and F-Measure of each level by the best performing model (MLP). Likewise, an obtained average TP Rate of **0.534** is obtained.

**Table 3** Results of classifiers using Groove Radar Values

| Model | ACC (%) | KP | MAE | RMSE |
|---|---|---|---|---|
| MLP | 66.13 | 0.5643 | 0.1224 | 0.2723 |
| LR | 65.96 | 0.5631 | 0.1236 | 0.2544 |
| NB | 64.89 | 0.5482 | 0.1285 | 0.2622 |
| Knn (K = 10) | 60.28 | 0.4823 | 0.1406 | 0.2723 |

**Table 4** Results of the MLP per class. The TP Rate of detecting Level 17 and 19s is poor

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| Lvl 13 | 0.903 | 0.095 | 0.835 | 0.903 | 0.868 |
| Lvl 14 | 0.537 | 0.101 | 0.558 | 0.537 | 0.547 |
| Lvl 15 | 0.560 | 0.080 | 0.602 | 0.56 | 0.580 |
| Lvl 16 | 0.541 | 0.078 | 0.513 | 0.541 | 0.526 |
| Lvl 17 | 0.439 | 0.042 | 0.450 | 0.439 | 0.444 |
| Lvl 18 | 0.590 | 0.021 | 0.676 | 0.590 | 0.630 |
| Lvl 19 | 0.167 | 0.004 | 0.333 | 0.167 | 0.222 |
| Average TP Rate = 0.534 | | | | | |

It can be clearly observed that classifiers have a hard time correctly identifying a difficulty level (treated as a category in DDR) using just the Groove Radar Values alone. The difficulty may be due to the dataset being unbalanced, but further analysis of the generated confusion matrix revealed that most of the errors occur due to a difficulty in distinguishing difficulties that are just one level away from an entry's correct level.

As observed in Table 4 and Fig. 6, the accuracy and the performance of the classifications using a tenfold cross validation with the MLP, LR, NB, and Knn models, majority of them managed to classify most of the dataset in the correct category (TP Rate being the actual percentage of the chart being classified as it's correct difficulty level). There were, however, some difficulties encountered in truly separating each difficulty level. Likewise, all classification models, with the exception of the Logistic Regression, were unable to classify Level 19s properly. This can be due to very small instances of what Level 19s are, and therefore a general consensus on what they are has not yet been reached. However, the Logistic Regression managed

(1) Confusion Matrix using MLP and 10-fold Cross Validation

| | Lvl13 | Lvl14 | Lvl15 | Lvl16 | Lvl17 | Lvl18 | Lvl19 |
|---|---|---|---|---|---|---|---|
| Lvl13 | 177 | 19 | 0 | 0 | 0 | 0 | 0 |
| Lvl14 | 32 | 58 | 16 | 1 | 1 | 0 | 0 |
| Lvl15 | 3 | 22 | 56 | 16 | 3 | 0 | 0 |
| Lvl16 | 0 | 4 | 19 | 40 | 9 | 2 | 0 |
| Lvl17 | 0 | 0 | 2 | 17 | 18 | 4 | 0 |
| Lvl18 | 0 | 1 | 0 | 4 | 9 | 23 | 2 |
| Lvl19 | 0 | 0 | 0 | 0 | 0 | 5 | 1 |

(2) Confusion Matrix using LR and 10-fold Cross Validation

| | Lvl13 | Lvl14 | Lvl15 | Lvl16 | Lvl17 | Lvl18 | Lvl19 |
|---|---|---|---|---|---|---|---|
| Lvl13 | 175 | 20 | 1 | 0 | 0 | 0 | 0 |
| Lvl14 | 30 | 56 | 22 | 0 | 0 | 0 | 0 |
| Lvl15 | 1 | 23 | 57 | 18 | 1 | 0 | 0 |
| Lvl16 | 0 | 2 | 21 | 41 | 7 | 3 | 0 |
| Lvl17 | 0 | 0 | 3 | 13 | 16 | 9 | 0 |
| Lvl18 | 0 | 1 | 0 | 4 | 8 | 23 | 3 |
| Lvl19 | 0 | 0 | 0 | 0 | 0 | 2 | 4 |

(3) Confusion Matrix using NB and 10-fold Cross Validation

| | Lvl13 | Lvl14 | Lvl15 | Lvl16 | Lvl17 | Lvl18 | Lvl19 |
|---|---|---|---|---|---|---|---|
| Lvl13 | 175 | 21 | 0 | 0 | 0 | 0 | 0 |
| Lvl14 | 33 | 55 | 19 | 0 | 0 | 1 | 0 |
| Lvl15 | 4 | 18 | 58 | 15 | 4 | 1 | 0 |
| Lvl16 | 1 | 3 | 20 | 39 | 11 | 0 | 0 |
| Lvl17 | 0 | 0 | 3 | 12 | 19 | 7 | 0 |
| Lvl18 | 1 | 0 | 1 | 5 | 11 | 19 | 2 |
| Lvl19 | 0 | 0 | 0 | 0 | 0 | 5 | 1 |

(4) Confusion Matrix using Knn and 10-fold Cross Validation

| | Lvl13 | Lvl14 | Lvl15 | Lvl16 | Lvl17 | Lvl18 | Lvl19 |
|---|---|---|---|---|---|---|---|
| Lvl13 | 177 | 19 | 0 | 0 | 0 | 0 | 0 |
| Lvl14 | 47 | 37 | 23 | 1 | 0 | 0 | 0 |
| Lvl15 | 6 | 24 | 59 | 8 | 3 | 0 | 0 |
| Lvl16 | 1 | 2 | 33 | 34 | 4 | 0 | 0 |
| Lvl17 | 0 | 0 | 9 | 16 | 14 | 2 | 0 |
| Lvl18 | 1 | 0 | 1 | 7 | 11 | 19 | 0 |
| Lvl19 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |

**Fig. 6** Confusion matrices generated by MLP (1), LR (2), NB (3), and Knn (4)

to catch the very high difference that separated Level 19s from the rest. Still, it was evident that there was some difficulty in differentiating between 18 and 19s.

## 4 Pattern Derived Attributes

It can clearly be observed from Sect. 3 that despite the Groove Radar Values' presence, while giving a clear estimate of a chart's contents and possibly difficulty setting, is still insufficient as there may be unmeasured qualities within each chart that differentiates between difficulty levels more explicitly. As a result, each chart will be analyzed further and additional attributes will be added for each chart that is a result of the patterns existing within the aforementioned chart. The pattern derived attributes mainly focus on the patterns that are extracted from parsing the step charts using string sequencing [11]. The attributes are as follows:

1. *Pattern Frequency Count (PFC)* = This indicates the frequency of that particular pattern existing within that chart. Sometimes, a pattern's prevalence in the stepchart (particularly crossovers) may affect the difficulty level of the chart. This is merely calculated as the number of times the pattern was encountered.
2. *Pattern Power Speed (PPS)* = This attribute indicates the fastest BPM (Beats Per Minute) required to execute that particular pattern. Sometimes, patterns become difficult and physically draining to execute at fast speeds and may impact the difficulty of a chart.

The patterns being referred to is indicated in Sect. 2. The patterns considered are the following patterns: *Jumps (Consecutive)*, *Staircase*, *Candlelight*, *Crossovers*, *Drills*, *Jacks*, and *Step Jumps*. Other patterns (*Steps, Freeze*, *Shocks*, *and Gallops*) are no longer considered as *Shocks* are not very prevalent in step charts and *Steps, Freeze, and Gallops* are obtained through the Groove Radar Value. *Shocks*, due to the nature of having to jump in the middle of the pad or jump high to avoid stepping on any arrow will be counted as a *Step Jump* or *Consecutive Jump* instead. The aforementioned attributes will be generated through string sequencing as a result of translating. SM files as mentioned in Sect. 2.1. The procedure is further illustrated in Fig. 7.

The resulting procedure will provide each dataset entry with an additional 14 attributes, each denoting of the 7 indicated pattern's PFC and PPS values.

## 5 Results from Classifications with Pattern Derived Attributes

After applying the Pattern Derived Attributes, a notable increase in the TP Rate for classifying all levels was observed as indicated in Table 5 and the overall accuracy indicated in Table 6. It is observed that the average TP rate has increased from

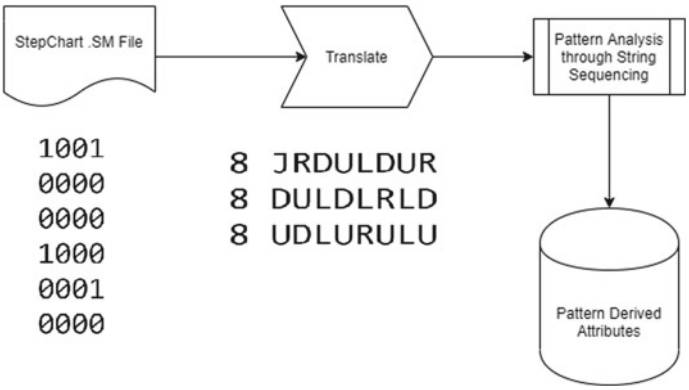



**Fig. 7** Process of obtaining pattern derived attributes from. SM files through string sequencing. Each measure of the. SM file are translated into a series of characters representing the chart

**Table 5** Result of adding the pattern derived attributes. TP rate for correctly classifying the levels has increased significantly from 0.534 to 0.683

| Class | TP rate | FP rate | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| Lvl 13 | 0.913 | 0.011 | 0.978 | 0.913 | 0.945 |
| Lvl 14 | 0.815 | 0.083 | 0.698 | 0.815 | 0.752 |
| Lvl 15 | 0.600 | 0.065 | 0.667 | 0.600 | 0.632 |
| Lvl 16 | 0.581 | 0.059 | 0.597 | 0.581 | 0.589 |
| Lvl 17 | 0.561 | 0.048 | 0.479 | 0.561 | 0.517 |
| Lvl 18 | 0.641 | 0.023 | 0.676 | 0.641 | 0.658 |
| Lvl 19 | 0.667 | 0.007 | 0.500 | 0.667 | 0.571 |
| Average TP Rate = 0.683 | | | | | |

**Table 6** Accuracies of the classifiers

| Model | ACC | KP | MAE | RMSE |
|---|---|---|---|---|
| NB | 74.82% | 0.6807 | 0.0845 | 0.2351 |
| MLP | 74.47% | 0.6727 | 0.0874 | 0.2351 |
| LR | 73.76% | 0.6651 | 0.0915 | 0.2274 |
| Knn (K = 10) | 70.74% | 0.6219 | 0.1018 | 0.2349 |

A noticeable improvement in the kappa statistic is observed

0.534 using only Groove Radar Values to **0.683** with the Pattern Derived Attributes included.

The Naïve-Bayes along with the addition of the Pattern Derived Attributes saw an increase in the overall accuracy and TP Rate along an increase in the Kappa values. This means that there are indeed some factors that can be derived in the

form of patterns sequenced as strings in determining a chart's difficulty level given a stepchart's designer on what corresponding rules make up a particular level for the game of Dance Dance Revolution.

## 6 Conclusion and Future Work

The study proposed a method that will be able to classify a chart's difficulty rating depending on attributes obtained from the patterns observed in the charts. Note that, although there are pre-computed values attributed to a chart, it has been shown that these values may be insufficient in properly determining the difficulty level. Likewise, a chart's final difficulty level upon deployment normally depends on what the specific chart designer's intuition is, and that normally has led to some chart being misrated. The approach shows promise that, even if there are no exact rules in determining a chart's level, a set of patterns can be derived that could possibly match a chart designer's intuition in identifying what that chart's level is. Note that this entirely depends on previous examples that exist as a result of charts that have been developed by the said designer.

This approach was designed with previous knowledge of existing patterns that exist in a specific rhythm game, however future work could involve developing a computational model that could automatically derive a common set of patterns that exist within charts of that particular game through string sequencing approaches and establish a set of rules in identifying what set of patterns can contribute to the chart's difficulty level in that particular game. Likewise, being able to derive other extrinsic values such as the distance between certain buttons, physical stamina needed, and mobility required to execute a given pattern taking into consideration a rhythm game's mechanics can be considered in building a more comprehensive model.

## References

1. Hoysniemi J (2006) International survey on the dance dance revolution game. Comput Entertainment (CIE)
2. Donahue C, Lipton Z, McAuley J Dance dance convolution. In: ICML'17 Proceedings of the 34th International Conference on Machine Learning, vol 70, pp 1039–1048
3. Machine learns to choreograph. https://www.onartificialintelligence.com/articles/10810/machine-learns-to-choreograph. Accessed on 20 June 2020
4. Tsujino Y, Yamanishi R (2018) Dance dance gradation: a generation of fine-tuned dance charts. In: 24th IFIP World Computer Congress, WCC 2018
5. DDR Songs with Re-Ratings. https://dancedancerevolution.fandom.com/wiki/Category:Songs_with_Reratings. Accessed on 13 June 2020
6. Yang L (2010) Modeling player performance in rhythm games. In: SIGGRAPH Asia 2010, Seoul, South Korea, 15–18 Dec 2010
7. Lee D (2016) Basic patterns that you need to know. https://ddrcommunity.com/basic-patterns-that-you-need-to-know/. Accessed on 13 June 2020

8. Stepmania. https://www.stepmania.com/. Accessed on 22 June 2020
9. Groove Radar 2019. https://dancedancerevolution.fandom.com/wiki/Groove_Radar. Accessed on 20 June 2020
10. Eibe F, Hall M, Witten I (2016) The WEKA workbench. Online appendix for "data mining: practical machine learning tools and techniques, 4th edn. Morgan Kaufmann
11. Crochemore M, Perrin D (1988) Pattern matching in strings. In: Cantoni V, Di Gesù V, Levialdi S (eds) Image analysis and processing II. Springer, Boston, MA

# Nasheed Song Classification by Fuzzy Soft-Set Approach

**Rabiei Mamat, Ahmad Shukri Mohd Noor, and Mustafa Mat Deris**

**Abstract** Classification of genres is among the important tasks of musical knowledge discovery. It may affect the accuracy of finding results or reducing the processing time when looking for a certain musical genre in an internet context. While the genre classification scheme looks very promising for western genres, the genre of non-western still has no space in genre retrievals, especially in identifying nasheed song. Therefore, a research has been carried out to select the best features to describe nasheed genre and creates a classifier using the selected features to classify nasheed. The features selection technique and the classifier were built based on the theory of fuzzy-soft set that have enough parameters to handle uncertainties in data. The result show that the built classifier using the selected features accurately can identify the nasheed genre up to 90%.

**Keywords** Soft-set · Fuzzy soft-set · Classification · Classifier · Nasheed Song

## 1 Introduction

Classification is one of the data mining operations used to assign an object to one of a several predefined category. In music information retrieval (MIR), it can be used to classify the song into the certain groups. This method, or better known as genre classification, is a daunting task as it deals with multi dimensional data with an overwhelm-

This research is suppported by RMIC, UMT.

R. Mamat · A. S. M. Noor (✉)
Universiti Malaysia Terengganu, Kuala Nerus, Terengganu 21300, Malaysia
e-mail: ashukri@umt.edu.my

R. Mamat
e-mail: rab@umt.edu.my

M. M. Deris
Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johore 86400, Malaysia
e-mail: mmustafa@uthm.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_18

ing number of features. Although the cost of storage is decrease, using the whole set of features will effects the processing speed [1–3]. At the same time, it may be difficult to interpret the outcome and reduced the accuracy of the classification result.

Unfortunately, the current MIR only supports the classification of genres related to western culture. The reason for all this is that research into genres such as jazz, classical, country and blues has grown rapidly due to the global researcher's attention to this type of genre. At the same time, the technical definition of the genre itself has become increasingly clear. However, there are some studies that take local folk songs, such as Malay traditional songs [8, 9], as well as Chinese traditional songs [6], as their case study and consideration.

This paper discussed the approaches and techniques used to develop a nasheed song classifier based on a fuzzy-soft set theory. The rest of the paper will be organized as follows. Section 2 explains the related research work on the subject in discussion. Section 3 provides an overview of the fuzzy-soft set theory. Section 4 sets out the modeling process. Next, the result and discussion are presented in Section 5 followed by the conclusions in Section 6.

## 2 Related Works

### *2.1 Nasheed Song*

Nasheed is an Arab branch of art and culture. It's a directing poem anda message of thanks to the prophet. It also covers inspirational stories and a message of thanks to Allah. Nasheed is a voice that is usually performed as a cappella. Yet often it's also accompanied by percussion instruments. Nasheed has become famous in the Islamic world.

There is a major change in nasheed today. Many of the modern nasheed artists are non-Arab and sing in different languages. Nasheed can be easily found in English, Malay, Urdu or Turkish. The new generation of nasheed artists uses a wide range of musical instruments in their art. They often introduce western influences, such as rock or pop, into the nasheed music, while retaining the basic features of the nasheed songs. This makes it difficult to identify the nasheed songs and maybe group them into the wrong genres.

### *2.2 Non-western Genre Classification*

According to [4], the successfulness of automatic song classification system depends to features that representing the music surface and rhythmic structure of audio signals. Thus, the first step to identifying a genre of a song is identifying the features of the genre itself or better known as feature selection which could differentiate them

from others. An assortment of method and approach was developed for non-western genre classification. For example, an experiments on the Indian music classifier have been carried out by [5]. In the experiment, two features: spectral shape and perceptual are evaluated using two classifiers: Gausian Mixture and k-Nearest Neighbor. They also found that some other musical features such as mel-frequency cepstral coefficient and Spectral centroid are interesting features to characterized Indian music. Li et al. [6] implemented the classification of Chinese folk songs by taken the temporal characteristics features and used Gaussian Mixture Model (GMM) and Restricted Boltzmann Machine (RBM) to calculate the label sequence. The findings of the experiments show that the solution suggested can reach a maximum accuracy of 84.71%, which beats the other classifiers used in the studies. Lashari et al. [7] used 2 features: perception based and—frequency cepstral coefficients (MFCCs) to classify traditional Pakistani musical instruments. For the classification task, they developed 2 classifiers that is based on the soft-set comparison table and fuzzy-soft set similarity measurement. Results show that both classifiers can perform well on numerical data. However, the accuracy of the soft set based classifier accuracy is higher. Senan et al. suggested the used of Rough Set Theory for feature selection to the problem of Traditional Malay musical instruments sounds. For the purpose of the feature selection, the introduced technique is then applied for feature ranking and attributes reduction. The accuracy rate of the selected features is measures using Support Vector Machine, Neural-Net and Naive-Bayes which implemented in Weka [8]. By the same approach, Senan et al. [9] also suggested the used of Soft-Set theory to solve the same problem.

## *2.3 Uncertainties Management*

Multi dimensional data with an overwhelming number of features definitely have higher uncertainties and difficult to understand. Molodtsov has introduced a new mathematical tool which claims to have sufficient parameterized methods to overcome uncertainties. The tool known as the soft-set theory [10] uses parametrization sets as its key to a problem-solving approach. It would make soft-set theory very easy and simpler to implement. However, in the theoretical and practical researches of soft sets, the situations are usually very complex. To this, Maji et al. [11] introduced the fuzzy-soft set principals which is influenced by the idea of fuzzy set. Majumdar and Samanta [12, 13] presented the generalised fuzzy soft sets and similarity measure between two fuzzy soft set. Cagman et al. [14] continue the fuzzy soft set with the idea of fuzzy set aggregation for decision-making problem. Fuzzy soft-set theory has been used in [15] to solve the problem of decision making by embedding rough-set approximation into the theory. Kalayathankal and Singh [16] applied fuzzy-soft set as tools for fuzzy analysis to simulate the unknown relations between a set of meteorological and hydrological parameters. By this, they try to predict and activating the flood alarm system when required. Handaga and Deris [17] introduced a method called fuzzy soft set classifier that used fuzzy-soft set in a field of text categorization using the idea of fuzzy c-means clustering. Previously, in [18], use the method from the Roy-Maji approach [11, 15] for medical decision making using the Cleveland dataset.

## 3 Overview of the Fuzzy-Soft Set Theory

As part of the soft-set theory extensions, the fundamental principle of soft-set and fuzzy-soft set theory are re-explained in this section to illustrate their variations and resemblance. See especially [10–12] for further details and background.

Let $U$ be the universe set and $E$ the set of all possible parameters under consideration with respect to $U$. Usually, parameters are attributes, characteristics, or properties of objects in $U$.

**Definition 1** [10] A pair $(F, A)$ is called a soft-set over $U$, where $A \subseteq E$ and $F$ is a mapping given by $F : E \to P(U)$.

Obviously, a soft set over $U$ is refered to any subset of $U$ parameterized by $E$. For a soft-set $(F, A)$, $A$ is the parameter set where $A \subseteq E$. For any $\alpha \in A$, $F(\alpha)$ can be regarded as a collection of approximate elements of $e$. Therefore, a soft-set across the universe $U$ can be characterized by a series of key-value pairs

$$(F, A) = \left\{\left(\alpha, F(\alpha)\right) : \alpha \in A, F(\alpha) \in P(U)\right\} \tag{1}$$

**Example 1** Let $U = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8, o_9\}$ is a set of marketing officer shortlisted described by their social media skills $E = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$. Each soft-skills $s_i \in E$ is respectively stand for facebook, twitter, instagram, linkedin, blog, whatsap and telegram. Suppose that each candidate has the skills as follows:
$o_1 = \{s_1\}, o_2 = \{s_1, s_4, s_5\}, o_3 = \{s_2, s_4\}, o_4 = \{s_1, s_4\}, o_5 = \{s_1, s_4, s_5\}, o_6 = \{s_3, s_5, s_6, s_7\}$, $o_7 = \{s_5\}$, $o_8 = \{s_2, s_4, s_5\}$ and $o_9 = \{s_2, s_3, s_5, s_5, s_7\}$.

Therefore, defining a soft-set $(F, E)$ as a subset of the universe $U$ parameterized by $s_i \in E$ will returned a list of the estimated description of an object which can be viewed as follows:

$$(FE) = \left\{F(o_1), F(o_2), F(o_3), F(o_4), F(o_5), F(o_6), F(o_7)\right\}$$

where;
$F(o_1) = \{s_1, s_2, s_4, s_5\}$,$F(o_2) = \{s_3, s_8, s_9\}$,$F(o_3) = \{s_6, s_9\}$,
$F(o_4) = \{s_2, s_3, s_4, s_5, s_8\}$,$F(o_5 = \{s_2, s_5, s_6, s_7, s_8, s_9\}$,$F(o_6) = \{s_6, s_9\}$ and
$F(o_7) = \{s_6, s_9\}$.

**Definition 2** [14] A fuzzy set $X$ across the universe $U$ is a pair $(U, m)$ where $m$ is a membership function such that $m : U \to [0, 1]$. The function $m = \mu_x$ is called the membership function of the fuzzy set $X = (U, \mu_x)$ where $\mu_x(x)$ is the membership grade for each $x \in U$.

Thus, a fuzzy set $X$ across the universe $U$ can be defined as follows

$$X = \left\{\left(\mu_x(u)/u\right) : u \in U, \mu_x(x) \in [0, 1]\right\} \tag{2}$$

Note that the set of all the fuzzy sets over $U$ will be denoted by $\widetilde{F}(U)$.

**Definition 3** [11, 14] Let $A_i \subseteq E$. A pair $(\widetilde{F}_i, A_i)$ is called fuzzy-soft set over $U$ where $\widetilde{F}_i$ is a mapping given by $\widetilde{F}_i : A_i \to \widetilde{F}(U)$.

$\widetilde{F}(a)$ is called fuzzy approximation function of the fuzzy-soft set $(\widetilde{F}, A)$, and the value $\widetilde{F}(a)$ is a set called $x - element$ of the fuzzy-soft set for all $x \in A$.

**Example 2** Let $U = \{o_1, o_2, o_3, o_4, o_5\}$ be a universal set and $E = \{p_1, p_2, p_3, p_4\}$ be a set of parameters. if $A = \{p_1, p_2, p_3, p_4\} \subseteq E$, $\widetilde{F}(p_1) = \left\{\frac{0.9}{o_2}, \frac{0.5}{o_4}\right\}$, $\widetilde{F}(p_2) = U$ and $\widetilde{F}(p_4) = \left\{\frac{0.2}{o_1}, \frac{0.4}{o_3}, \frac{0.8}{o_5}\right\}$, then a fuzzy-soft set $(F, A)$ can be written by $(\widetilde{F}, A) = \left\{\left(x_1, \left\{\frac{0.9}{o_2}, \frac{0.5}{o_4}\right\}\right), (x_2, U), \left(x_4, \left\{\frac{0.2}{o_1}, \frac{0.4}{o_3}, \frac{0.8}{o_5}\right\}\right)\right\}$.

**Definition 4** [17] Let $F_A$ be a fuzzy-soft set over $U$. The cardinal set of $F_A$ which is denoted by $\zeta_{F_A}$ is defined by

$$\zeta_{F_A} = \left\{\mu_{\zeta_{F_A}}(x)/x : x \in E\right\} \tag{3}$$

is a fuzzy set over $E$. The membership function $\mu_{\zeta_{F_A}}$ of $\zeta_{F_A}$ is defined by $\mu_{\zeta_{F_A}} : E \to [0, 1]$ where

$$\mu_{\zeta_{F_A}}(x) = \frac{|\gamma_A(x)|}{|U|} \tag{4}$$

where $|U|$ is the absolute value of universe U, and $|\gamma_A(x)|$ is the scalar cardinality of fuzzy set $\gamma_A(x)$.

**Definition 5** [12, 13] Similarity between two fuzzy-soft set $(\widetilde{F}, E)$ and $(\widetilde{G}, E)$ denoted by $S(\widetilde{F}, \widetilde{G})$ or $S_{\widetilde{F},\widetilde{G}}$ is defined as follows:

$$S(\widetilde{F}, \widetilde{G}) = S_{\widetilde{F},\widetilde{G}} = \frac{\sum_{i=1}^{n}\{\widetilde{F}(e_i) \cdot \widetilde{G}(e_i)\}}{\sum_{i=1}^{n}\{(\widetilde{F}(e_i))^2 \vee (\widetilde{G}(e_i))^2\}} \tag{5}$$

## 4 Modelling Process

### *4.1 The Algorithm*

The algorithm for classifier used in this research is taken from [17] with some adjustment. The details of formula is explained in section 3. The full algorithm is as follows (Table 1):

Based on the algorithm, the classifier has been build using the python programming language.

### *4.2 The Data*

This research used 100 contemporary nasheed songs as sample data. Each song then separated into 3 units which representing the beginning of a song, a middle of a song

**Table 1** Fuzzy-soft set classifier for Nasheed Song

| Algorithm1 : Fuzzy-Soft set Classifier<br>Input:trainData, testData |
|---|
| 1: $(F, E) \leftarrow fuzzification(trainData)$ |
| 2: $(\widetilde{F}, E) \leftarrow cardinalSet((F, E))$ |
| 3: $(G, E) \leftarrow fuzzification(testData)$ |
| 4: $(\widetilde{G}, E) \leftarrow cardinalSet((G, E))$ |
| 5: $S(\widetilde{G}, \widetilde{F})$ |
| 6: $threshold \leftarrow setThreshold()$ |
| 7: If $S(\widetilde{G}, \widetilde{F}) \geq threshold$ |
| 8: $nasheed \leftarrow (\widetilde{G}, E)$ |
| 9: EndIf |

and end of a song which produces 300 units of a sample. The size of each unit is 512 kb with 16kHz sample rate.

Two features have been recognized to be used in this research ie: Area Method of Moments with 10 Attributes and Area Method of Moments of Mel Frequency Cepstral Coefficient (MFCC) with 10 Attributes. Data extraction is carried-out using jAudio-1.0.4 by the dot wav input and produced dot ARFF output. Class labelling process is made by a java program where after labelling, dot ARFF is converted into dot file in the form of information system $S = (U, A \cup \{d\}, v, f)$.

### *4.3 Evaluation and Validation*

Evaluation process is carried out using cross-validation approach. Fifty units of sample data are taken randomly as a nasheed training dataset while another random 50 units is used as a testing data. The accuracy of the classification is measure using the following formula:

$$Accuracy = \frac{Number\ of\ \ Correct\ Prediction}{Total\ Number\ of\ \ Prediction} \tag{6}$$

As a comparison to the other genre, the world genre data is taken from http://marsyas.info is used as the testing data. The dataset used is the same as applied in [4].

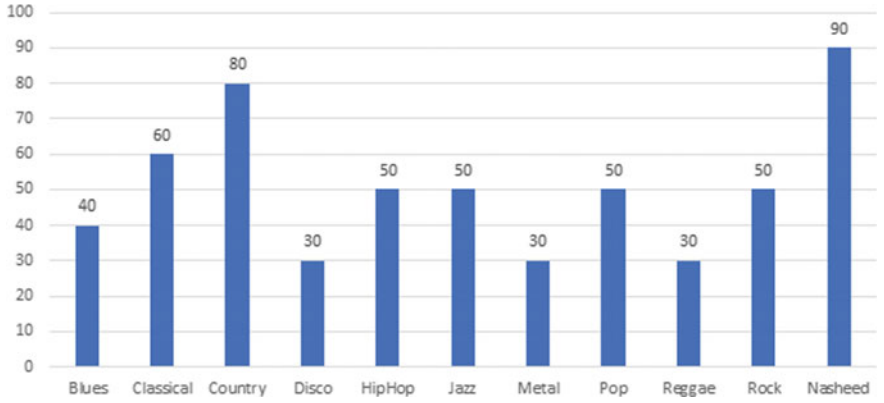



**Fig. 1** Classification Result Using Nasheed Song Classifier

## 5 Result and Discussion

The result of classification using Nasheed song classifier is shows in Fig. 1. In summary, it can be seen in the graph that 40% of blues genre test data, 60% of classical genre test data, 80% of classical genre test data, 30% of disco genre test data, 50% of hip-hop genre test data, 50% of jazz genre test data, 30% of metal genre test data, 50% of pop genre test data, 30% of reggae genre test data, 50% of rock and 90% nasheed genre test data are classified as nasheed songs correspondingly.

The accuracy of the classifier for nasheed genre which is up to 90% clearly shows that the selected features can be used by the fuzzy-soft set to differentiate the nasheed genre. This is also supported by other genre classification results that are below or equals to 50% accuracy such as blues, disco, hip-hop, jazz, metal, pop, reggae, and rock.

Unfortunately, the accuracy of two genres ie classic and country are still questionable. The accuracy of the classic and country which is 80% and 60% respectively may be due to the selected features used. However, this issue requires further investigation.

## 6 Conclusion

In this paper, the applicability of fuzzy-soft set theory for nasheed song classification is investigated. A classifier that is based on the fuzzy-soft set has been developed. A collection of 100 nasheed songs that already going thru the data cleaning process is used as the training data as well as the testing data. The extracted data only contains two features ie Area Method of Moments and Area Method of Moments of Mel

Frequency Cepstral Coefficient (MFCC) that have been recognized earlier. Beside that, a world genre collections is used as a comparison data.

The result show that the built classifier using the selected features accurately can identify the nasheed genre up to 90%. Thus, it explains the applicability of fuzzy-soft set in musical information retrieval (MIR).

## References

1. Chang K, Jang JSR, Iliopoulos CS (2010) Music genre classification via compressive sampling. In: Proceedings international society for music information retrieval. Amsterdam, The Netherlands, pp 387–392
2. Gjerdingen, RO, Perrott David (2008) Scanning the dial: the rapid recognition of music genres. J New Music Res 37(2):93–100
3. Mandel MI, Ellis DPW (2008) Multiple-instance learning for music information retrieval. In: Proceedings of the 9th international conference of music information retrieval. Drexel University, pp 577–582
4. Tzanetakis G, Essl G, Cook P (2002) Automatic musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(5):293–302
5. Jothilakshmi S, Kathiresan N (2012) Automatic music genre classification for Indian Music. In: International Conference on Software and Computer Applications (ICSCA 2012) IPCSIT, vol 41. IACSIT Press, Singapore
6. Li J, Luo J, Ding J, Zhao X, Xinyu Y (2018) Regional classification of Chinese folk songs based on CRF model. Multimedia tools and applications
7. Lashari SA, Ibrahim R, Senan N (2012) Performance comparison of musical instrument family classification using soft set. Int J Artif Intell Expert Syst (IJAE) 3(4)
8. Senan N, Ibrahim R, Nawi MN, Yanto ITR, Herawan T (2011) Rough set approach for attributes selection of traditional Malay musical instruments sounds classification, UCMA 2011. CCIS 151:509–525
9. Senan N, Ibrahim R, Nawi MN, Yanto ITR, Herawan T (2010) Soft set theory for feature selection of traditional Malay musical instrument sounds, ICICA 2010. LNCS 6377:253–260
10. Molodtsov D (1999) Soft set theory: first results. Comput Math Appl 37(4–5):19–31
11. Maji P, Biswas R, Roy A (2001) Fuzzy soft sets. J Fuzzy Math 9(3):589–602
12. Majumdar P, Samanta S (2011) Similarity measure of fuzzy-soft sets. Int J Adv Soft Comput 3(2)
13. Majumdar P, Samanta S (2010) Generalised fuzzy soft sets. J Comp Math App 59:1425–1432
14. Cagman N, Enginoglu S, Citak F (2011) Fuzzy soft set theory and its application. Iranian J Fuzzy Syst 8(3):137–147
15. Roy AR, Maji PK (2007) A fuzzy soft set theoretic approach to decision making problems. J Comput Appl Math 203(2):412–418
16. Kalayathankal SJ, Singh GS (2010) A fuzzy soft flood alarm model. Math Comput Simulation 80(5):887-893. ISSN 0378-4754
17. Bana H, Tutut H, Mat DM (2012) FSSC: an algorithm for classifying numerical data using fuzzy soft set theory. Int J Fuzzy Syst Appl
18. Kirişci Murat (2020) Medical decision making with respect to the fuzzy soft sets. J Interdisc Math 23(4):767–776. https://doi.org/10.1080/09720502.2020.1715577

# Hybrid SDN Deployment Using Machine Learning


**H. W. Siew, S. C. Tan, and C. K. Lee**

**Abstract** Software-Defined Networking (SDN) has attracted tremendous attention in recent years as the future communication network architecture. However, SDN deployment in legacy network will be progressively phased over a period, especially for larger network which consists of hundred or more nodes. Every migration (i.e. replacing or upgrading) of SDN-enabled nodes requires considerable optimization efforts in terms of cost of investment, network stability and performance gains. Hitherto literatures have proposed variety of static heuristic algorithms to compute the migration sequence of SDN-enabled nodes for multi-periods SDN deployment in legacy network. The aim of each computed migration sequence is aims to improve network performance gains with respect to address different constraints. However, the dynamicity of an unique network, such as traffic growth or topology change, cannot be comprehensively addressed using a static heuristic algorithm over the deployment duration. Machine learning (ML), on the other hand, has been proven successfully applied for various dynamic and non-linear problems in diverse domains. In this article, we summarize the generic workflow for ML in networking domain at first. Subsequently, we investigated the problem of SDN deployment in legacy network from the perspective of ML. We proposed a SDN deployment problem that formulated as Markov Decision Process and reinforcement learning techniques, such as Qlearning and SARSA, can be used to model for the problem.

**Keywords** Machine learning · Software-defined networking · Hybrid SDN deployment

H. W. Siew · S. C. Tan (✉)
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
e-mail: sctan1@mmu.edu.my

C. K. Lee
Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_19

## 1 Introduction

Machine learning (ML) applies widely in diverse domains, such as speech recognition, computer vision, and autonomous vehicle. The unprecedented power of ML enables a system to extract knowledge from quality data [1]. Unlike the traditional programming approach of tackling a problem, ML uncover hidden rules or patterns via the discovery process with data (a.k.a. the learning process). The learned pattern, i.e. model, is then used to answer unknown data of the particular problem. In general, there are four classes of problems can leverage on the techniques of ML, namely, classification, clustering, regression, and rule extraction [2]. In classification and regression problems, new input data is mapped respectively to discrete or continuous output value. Whereas, the goal of clustering problem is to divide data points into groups alike. On the contrary, rule extraction problems are inherently different from others which the objective is to establish statistical relationships in data. A different learning paradigm of ML is employed for each class of problems. Supervised learning uses labelled data to identify hidden behaviours within datasets. Commonly, supervised learning is used to create model of classification and regression problems, in which, to classify discrete or to predict continuous output value. However, labelled data are not always available for all problems. Unsupervised learning, on the other hand, utilizes unlabelled data to train and learn knowledge from datasets. Ultimately, unsupervised learning targets to discriminate groups in the data, and this approach best suits clustering problems. In contrast, reinforcement learning is a machine learning paradigm which constantly learns through interactions (a series of actions) with the environment and observe the result to adjust its strategy automatically [3]. Consequently, the agent aims to extract optimal rules or policy for the problem.

Recently, ML has regained attraction in communications and networking domain to improve how networking problems are addressing today [4]. Communication networks are growing shockingly complex with wide spectrum of applications. Network operation and management remains tedious and error prone with human factor involved [5]. The diversity and complexity of communication networks has made designing scalable network solutions difficult. Often, solutions are built for network scenarios specific to its particular, such as type of applications, user demand and topology. Nonetheless, it is challenging to model an accurate representation of a complex network behaviours, for instance, loads pattern in Content Distribution Network (CDN) [6]. Furthermore, the dynamics of network inhibits developing efficient algorithms to cater different scenarios across networks. Therefore, there is a raising demand for cognitive management and operation in today networks [7]. Network operators are growing interest to build a highly resilient autonomic network as proposed in [8]. In which, an autonomic network comprises of self-configuration, self-healing, self-optimization and self-protection characteristics. Future networks is likely to operate autonomously by monitoring its own state and the environment together with their complex configuration. And, techniques of ML serve as a tool to facilitate decision making and network automation [4].

The idea of incorporating intelligence into network management and operation has been discussed for years in research community. However, such system has not been deployed or developed yet in existing networks [9]. One of the biggest challenges is that exiting network architecture is inherently distributed in nature [10]. Switches and router have only limited view and control over the whole network. For instance, legacy network devices are still restricted by vendor specific commands and functionalities, and It is extremely difficult to orchestrate such devices in heterogeneous network environment. Data availability and processing poses another challenge regards the deployment of autonomic system in existing networks [11]. Questions often arise where and what data can be collected from existing network. In addition, the training process of ML techniques with data requires tremendous computing and storage resources. Nevertheless, the advancement of recent technology has lowered the barrier for ML adoption in networking. Software-Defined Networking (SDN) [12], for example, decouples network controls (control plane) from its forwarding (data plane). The separation of control and data planes offers programmability through centralized controller. The programmability enables external software (e.g. ML applications) to define network behaviours with global network view. The prevalent of cloud computing nowadays alleviates the obstacle of demanding computing resources for ML techniques [13]. Even more, the availability of Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) in cloud accelerates the model training process of ML.

SDN promotes the applicability of ML for networking in which centralized controller offers global network view on top of programmability. However, the adoption of SDN in legacy network encountering several challenges [14], such as financial and technological constraints. To illustrate, budget consideration prohibits larger network operator to migrate hundreds of nodes to SDN-capable devices at once. Hence, SDN deployment in legacy network is likely to migrate nodes spanning over a period. The coexistence and interoperability of both legacy and SDN-capable devices in a network forms a hybrid SDN (hSDN) [15]. The deployment of hSDN involves legacy nodes selection for SDN-capable device migration in each period. Network operators are eager to understand with limited resources which legacy node and when it should be migrated. Ultimately, the deployment of hSDN can reap most benefits in term of network performance gain. Many studies [16–20] have been contributed to propose algorithms in seeking optimal migration sequence (order of nodes to migrate) of hybrid SDN deployment. In essence, an algorithm captures a snapshot of network traffic condition and compute the maximized performance gain with respect to different constrains, such as budget and link capacity. However, this approach underestimates the volatile of a network where traffic demand fluctuates, and network topology is subject to change over time. Hence, these algorithms fail to adapt the dynamicity of a network in the course of hybrid SDN deployment over months or years.

Literature of hybrid SDN deployment in legacy network exposes a research gap, in which, the dynamicity of a network has not been properly addressed in the previously studies. On the contrary, ML offers inherent adaptability with data learning process which caters dynamicity. Therefore, we are interested to investigate the applicability

of ML in the aspect of SDN node deployment. In this article, we aim to provide our insights and a new perspective to apply ML technique with hSDN deployment in legacy network. The following summarizes the contributions of this.

Generic workflow of ML in networking is summarized here to provide a basic practical guideline for applying ML in networking.

We expand the generic workflow of ML and illustrate the feasibility to apply ML technique for hSDN deployment problem in legacy network.

The following of this articles, related work is discussed in Sect. 2. In Sect. 3, we summarize the generic workflow of ML in networking. We expand the generic workflow in Sect. 4 with respect to SDN node migration in legacy network. Lastly, we conclude the article in Sect. 5.

## 2 Related Works

Large body of literature [16, 17, 19, 21–23] have been proposed to offer static heuristic algorithms computing the node migration sequence for hSDN deployment in legacy network. On one hand, these works consider the network topology remains unchanged throughout the deployment of hSDN. On the other, these proposed algorithms do not capture the possible growth of traffic in a network. Generally, the assumptions of both, unchanged topology and stagnant traffic, are unrealistic for a deployment timeframe possibly in years. Although [24] take into account of multiple traffic matrices in its computation of migration sequence, it considers only the past traffic fluctuations which does not sufficiently represent the dynamic (i.e. network topology change) a network may have after the deployment started. ML has been proven applicable to solve various problems in vast domains including networking. In networking, ML has play an significant role in traffic prediction [25] to forecast future traffic, traffic classification [26] to facilitate network operations and planning, traffic routing [27] and congestion control [28] to optimize resource utilization. Moreover, ML can also work in network management activities such as QoS management [29], fault management [30] and network security [31]. Nonetheless, SDN deployment planning has not been extensively discussed in the perspective of ML techniques [32]. Hence, instead of the static heuristic approach, we examine the feasibility of applying ML techniques in the problem of hSDN deployment in legacy network here.

## 3 Generic Workflow for Machine Learning in Networking

Figure 1 illustrates a generic workflow to apply machine learning in various domains including fields of networking [13]. In summary, the workflow consists of multiple steps, namely problem formulation, data collection, feature engineering, model construction and evaluation. Each step in the workflow is not independent but strongly interrelated between problem, training data and learning paradigm [11]. For instance,


**Fig. 1** The generic workflow of machine learning in networking

the result of a ML model depends largely on the collection and preprocessing of available data. In this section, we review each step in details in order to properly develop machine learning applications for networking related problems.

**Problem Formulation**: There are many different possible approaches to leverage ML for a networking problem. However, the process of training a ML model often requires huge amount of resources (e.g. time and investment). Therefore, it is utmost important to formulate correctly a networking problem at hand in the first step. Otherwise, an ill-formed problem will end at unsatisfactory performance as a result. In this step, a well formulated problem can be categorized into one of the ML paradigms, such as supervised, unsupervised or reinforcement learning. This helps to determine what kind of data are required for collection, and what learning algorithms to choose from for model construction. For example, a problem cannot be formulated as supervised learning if there is lacking label data for model training.

**Data Collection**: Various ML techniques share one common requirement which large amount of unbiased representative data is necessary to build an effective model for a designated problem in networking. According to the needs, network data can be monitored and recorded from different network layers, for instance Simple Network Management Protocol (SNMP) [11]. It is important to note that representative data vary from one problem to another even in the same domain. For instance, traffic prediction and traffic classification require different details of network data. Typically, data collection in the context of ML for networking is accomplished in two phases, offline and online. A sizable amount of historical data is gathered in the offline phase for model training and testing purpose. In online phase, real-time network data (e.g. performance information and network state) are collected and feed as input for the model retraining or used as a feedback to the model.

**Feature Engineering**: Every problem in networking is characterized by a number of factors. However, only few of defining factors (i.e. features) has the significant impact on the targeted network problem. Broadly, these features are categorized by its granularity level, for instance, connection-level, flow-level, and packet-level features [11]. The extraction of defining features in this step is crucial to unleash the pattern in data via different ML paradigms. The goal of this step attempts to analyze historical data and extract the effective features for model construction in next. Nonetheless, network data collected are often noisy or incomplete. Therefore, it is necessary to clean up data by going through a preprocessing phase prior to feature extraction. Feature extraction can be difficult which requires to have a thorough understanding regards the target problem with domain-specific knowledge [6]. Deep learning, on

the other hand, can ease to automate feature extraction in some of the problems in networking.

**Model Construction**: In model construction step, a suitable learning algorithm is chosen in accordance to the characteristics (e.g. problem category, size of dataset and etc.) of the target network problem defined. Prior to start training the selected model, the collected historical dataset is divided into training, validation and test datasets. It is important that all training, validation and test datasets are independent but follow the same probability distribution [11]. This prevents the generalization of training outcome which leads to model overfitting or under-fitting. Along the way of training selected model, training dataset also helps for hyper-parameter tuning in the offline phase. The process of parameters tuning involves finding acceptable parameters for building the selected model. While training dataset used for training, test dataset is used to evaluate the accuracy of the trained model. It is possible that this step is repeated until satisfied result is obtained.

**Model Validation**: Validation is an essential step for ML workflow in order to assess the performance of the trained model and how well it would generalize to new data. K-fold cross validation is often used to validate the accuracy of a model in overall. The result of validation offers insights on how to further optimize the model.

Furthermore, sources of error can be identified through model validation to determine if the model or feature are appropriate or data are representative enough for the target problem. If necessary previous steps can be re-visited according to the error sources discovered in the validation step.

## 4 Machine Learning for hSDN Deployment

The usages of the Internet have evolved over the past decades which increasingly demand network infrastructure to handle dynamic nature of future network operations and applications. For this reason, SDN has gained tremendous attention for the past few years as a next generation architecture of communication network. Among all, the most distinctive features of SDN are listed as follow,

1. Separation of network control (control plane) from forwarding (data plane)
2. A centralized controller with global network view
3. Network programmability by the external applications.

In short, SDN offers flexible configurability by external software, and allows network to be dynamically optimized in conjunction with the global network status. The benefit of SDN over its advantage of network controllability is tempting to network operators. However, full SDN deployment in an existing legacy network encounters challenges in organizational, economical, and technical aspects [14]. Nonetheless, literature has suggested that the benefit of SDN can be realized without the need to fully deploy SDN nodes in a legacy network [16, 20]. In most cases, network operators tend to gradually deploy hSDN in legacy networks which span

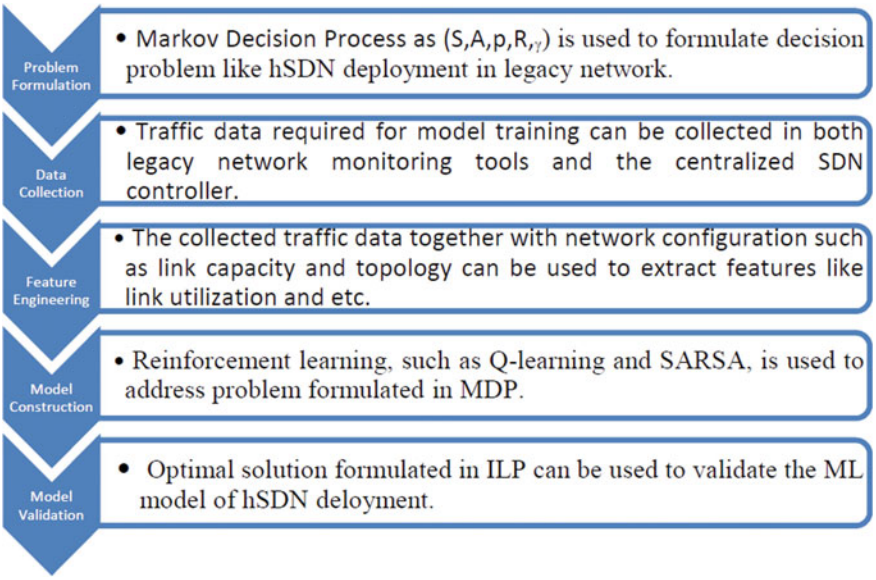



**Fig. 2** ML framework for hSDN deployment in legacy network

across multiple periods over months or years. The gradual deployment of hSDN in legacy network mainly results from the budget and technical considerations, especially for large network which consists of hundred or more nodes. In each deployment period, a number of nodes are selected to be migrated (e.g. replace or upgrade) to SDN enable devices in line with the budget available and other constraints (e.g. link capacity). Unlike the deployment of full SDN, hSDN deployment requires detail planning in selection of nodes to be migrated. Network operators are keen to understand which node at when should be migrated to maximize the return of investment. In this section, we apply the previously discussed workflow of ML for the problem of SDN deployment in legacy network as summarized in Fig. 2.

**Problem Formulation**: The deployment of hSDN in legacy network has been formulated as an optimization problem in literature. In which, hSDN deployment aims to optimize particular network performance matrix, e.g. maximum link utilization (MLU) or alternative paths, with respect to a number of constraints such as link capacity, budget and etc. Moreover, the hSDN deployment problem is an offline migration operation during network planning phase. During the planning process, decisions have to be made which legacy node at when should be migrated for hSDN deployment. Consequently, the decision making nature of hSDN deployment problem can be cast as Markov Decision Process (MDP) as illustrated in Fig. 3. In general, a tuple $(S, A, p, R, \gamma)$ is used to represent a MDP where $S$ denotes a finite set of states, $A$ denotes a finite set of actions, $p$ denotes the probability of state transition from $s$ with action $a$ to $s'$, $R$ is the reward obtained after execution of action $a$, and discount factor $\gamma \in [0,1]$ represents the importance of future reward as compared to

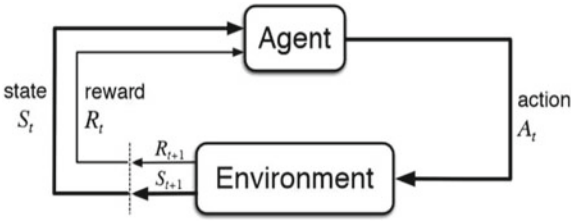



**Fig. 3** Markov decision process

the immediate reward. Simply put, the goal of a MDP is to maximize the total reward by finding an optimal policy $\pi *$, and policy $\pi$ is a function mapping from a state to an action.

**Data Collection**: Representative network data, for instance traffic matrix (TM), is necessary to properly train a ML model for hSDN deployment problem. Before hSDN deployment in legacy network, historical TMs can be acquired from network monitoring tools such as Cisco Net-Flow and IP Flow Information Export [11]. Once the deployment of hSDN started, network data can be aggregated from both legacy network monitoring tools and the centralized SDN controller. Among all, dynamic network data like TMs helps to understand the fluctuation of traffic demand in a network, and other data such network topology, links capacity and links weight settings are required to understand the configuration of a network.

**Feature Engineering**: MDP requires states to be defined for an environment. In the context of hSDN deployment, the environment consists of both legacy and SDN-enable nodes in a network. With every action of deploying a SDN node, the environment changes (i.e. states) in return for network performance gain. Eventually, a series of SDN node placement needs to be determined in order to achieve the optimal performance improvement. A state, in that sense, should be efficiently represented by distinctive features that reflects the performance gain acquired through the placement of SDN node. Specifically, features for hSDN deployment problem include link utilization, MLU, number alternative paths and etc. These features can be extracted from collected data such as TM and network topology configuration.

**Model Construction**: Commonly, reinforcement learning (RL) is used to address problem formulated as an MDP. Reinforcement learning algorithms learn through trial and error by interacting with the environment. An agent sequentially makes decisions (actions) and observes the outcomes (rewards) in environment (states). Ultimately, an agent targets to achieve the optimal policy through adjusting its strategy in making decisions [3]. For a problem which transition probability and reward models are known, model-based reinforcement learning such as value-iteration or policy-iteration algorithm can be used to solve the problem. In most cases, however, transition probability and reward models are difficult to define in a dynamic environment like the hSDN deployment in legacy network. Thus, a model-free reinforcement learning such as Q learning or SARSA can be used for ML model construction.

**Model Validation**: Typically, hSDN deployment is formulated as integer linear programming (ILP) problem in literature. In which, the ILP of hSDN deployment maximizes the network performance gain with respect to constraints, such as budget and link capacity. Optimal solution of such ILP problem can be computed using an ILP solver. Therefore, an optimal solution serves as a good candidate for benchmarking the RL model developed for hSDN deployment. However, ILP solver can take significant amount of time to compute an optimal solution for a large network. Heuristic algorithms developed in literature can then be used to approximate a sub-optimal solution for validating the RL model.

## 5 Conclusion

In summary, the dynamic nature of future communication network demands flexible network controllability. SDN has attracted tremendous attention recent years to provide such controllability via standardization and centralized controllers. However, full SDN deployment in legacy network is not always possible especially for large network due to various considerations. Large body of literature had proposed static approaches to offer gradual deployment of SDN in legacy network. However, networks' dynamicity, such as growth of traffic or change in topology, has not been properly addressed in the past. Machine learning, on the other hand, has been applied in various aspects of networking other than SDN deployment in legacy network. Techniques of machine learning have shown great success in revealing pattern with dynamic data. Thereby, we summarized the generic workflow of ML in networking, and we proposed to formulate hSDN deployment in legacy network using Markov Decision Process. In additional, technique of reinforcement learning, such as Q-learning or SARSA, can be used for model construction and validation. Moreover, further work is necessary to develop and evaluate the proposed approach here for hSDN deployment in legacy network.

## References

1. Brill E, Lin J, Banko M, Dumais S, Ng A (2002) Data-intensive question answering. In: Proceedings of the TREC-10 conference, pp 183–189
2. Practical machine learning problems. https://machinelearningmastery.com/practical-machine-learning-problems/. Accessed 11 Nov 2019
3. Luong NC et al (2019) Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun Surv Tutor 21(4):3133–3174
4. Chemouil P et al (2019) Artificial intelligence and machine learning for networking and communications. IEEE J Sel Areas Commun 37(6):1185–1191
5. Mahmoud QH (2007) Cognitive networks: towards self-aware networks. Wiley, Hoboken
6. JiangJ, Sekar V, Zhang H, Milner H, Shepherd D, Stoica I (2016) CFA: a practical prediction system for video QoE optimization

7. Ramming JC, Wroclawski JT, Clark DD, Partridge C (2015) A knowledge plane for the internet. In: Proceedings of the 2007 conference on applications, technologies, architectures, and protocols for computer communication, pp 3–10
8. White RS, Hanson EJ, Whalley I, Chess MD, Kephart JO (2004, October) An architectural approach to autonomic computing ('An architectural blueprint for autonomic computing'). In: International conference on autonomic computing, pp 2–9
9. Mestres A et al (2017) Knowledge-defined networking. Comput Commun Rev 47(3):1–10
10. Ayoubi S et al (2018) Machine learning for cognitive network management. IEEE Commun Mag 56(January):158–165
11. Boutaba R et al (2018) A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. J Internet Serv Appl 9(1)
12. Chen J, Zheng X, Rong C (2015) Survey on software-defined networking. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 9106, no 1, pp 115–124
13. Wang M, Cui Y, Wang X, Xiao S, Jiang J (2018) Machine learning for networking: workflow, advances and opportunities. IEEE Netw 32(2):92–99
14. Vissicchio S, Vanbever L, Bonaventure O (2014) Opportunities and research challenges of hybrid software defined networks. ACM SIGCOMM Comput Commun Rev 44(2):70–75
15. Amin R, Reisslein M, Shah N (2018) Hybrid SDN networks: a survey of existing approaches. IEEE Commun Surv Tutorials 20(4):3259–3306
16. Poularakis K, Iosifidis G, Smaragdakis G, Tassiulas L (2019) Optimizing gradual SDN upgrades in ISP networks. IEEE/ACM Trans Netw 27(1):288–301
17. Poularakis K, Iosifidis G, Smaragdakis G, Tassiulas L (2017) One step at a time: optimizing SDN upgrades in ISP networks. In: Proceedings—IEEE INFOCOM, pp 1–9
18. Guo Y, Wang Z, Yin X, Shi X, Wu J, Zhang H (2016) Incremental deployment for traffic engineering in hybrid SDN network. In: 2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC 2015)
19. Xu H, Li XY, Huang L, Deng H, Huang H, Wang H (2017) Incremental deployment and throughput maximization routing for a hybrid SDN. IEEE/ACM Trans Netw 25(3):1861–1875
20. Levin D, Canini M, Schmid S, Feldmann A (2013) Incremental SDN deployment in enterprise networks. In: Proceedings of the ACM SIGCOMM 2013 conference SIGCOMM—SIGCOMM'13, p 473
21. Tanha M, Sajjadi D, Ruby R, Pan J (2018) Traffic engineering enhancement by progressive migration to SDN. IEEE Commun Lett 22(3):438–441
22. Yuan T, Huang X, Ma M, Zhang P (2017) Migration to software-defined networks: the customers view. China Commun 14(10):1–11
23. Wang W, He W, Su J (2017) Boosting the Benefits of Hybrid SDN. In: Proceedings of the international conference on distributed computing systems, pp 2165–2170
24. Guo Y, Wang Z, Yin X, Shi X, Wu J (2017) Traffic engineering in hybrid SDN networks with multiple traffic matrices. Comput Netw 126:187–199
25. Poupart P et al (2016) Online flow size prediction for improved network routing In: Proceedings of the International Conference on Network Protocols ICNP, December 2016, pp 1–6
26. Wang R, Liu Y, Yang Y, Zhou X (2006) Solving the app-level classification problem of P2P traffic Via optimized support vector machines. In: Proceedings of the ISDA 2006 sixth international conference on intelligent systems design and applications, vol 2, pp 534–539
27. Hu T, Member S, Fei Y (2010) QELAR: a Machine-learning-based adaptive routing protocol for. IEEE Trans Mob Comput 9(6):796–809
28. Jayaraj A, Venkatesh T, Murthy CSR (2008) Loss classification in optical burst switching networks using machine learning techniques: improving the performance of TCP. IEEE J Sel Areas Commun 26(6):45–54
29. Demirbilek E, Gregoire JC (2017) Machine learning based reduced reference bitstream audio-visual quality prediction models for realtime communications. In: Proceedings of the IEEE international conference on multimedia and expo, vol 13, no 2, pp 571–576

30. Kumar Y, Farooq H, Imran A (2017) Fault prediction and reliability analysis in a real cellular network. In: 2017 13th International Conference on Wireless and Mobile Communications IWCMC 2017, pp 1090–1095
31. Cannady JD (1998) Artificial neural networks for misuse detection. In: Proceedings of the 21st national information systems security conference, pp 368–381
32. Wei SH, Chin TS, Binlun JN, Kwang LC, Kapsin R, Yusoff Z Machine learning as a means to adapt requirement changes for SDN deployment process in SDN migration. In: Advances in computational intelligence, pp 629–639.

# LED Lighting Assessment for High-Performance Stadium Illuminance


**Najmuddin Salmi bin Mat Nanyan, It Ee Lee, Gwo Chin Chung, and Duu Sheng Ong**

**Abstract** The usage of LED lighting has been commonly used in indoor and outdoor to facilitate energy-efficient. Because of its characteristics that relatively consume less energy compared to other traditional forms of lighting, it is considered the best alternative to be used as a lighting source. LED lighting also can be applied in stadium lighting applications. The lighting requirement for a stadium is exceptionally high. It requires many luminaires to be used to meet these specifications. The specifications consist of several parameters that inter-related that will affect the performance of lighting. Besides, a different type of view angle of a luminaire will give a different visual performance. Due to this challenge, it requires thorough work being done during the design process as well as during the installation process. The luminaires need to be aimed at proper aiming point coordinates to meet the specifications. Hence, a study on the characteristics of LED lighting needs to be done. In this paper, the objective is to model the output from LED lighting and study the effect of tilt angle and view angle of the luminaire on the visual performance of stadium lighting. A computational model of LED luminaire was developed using MATLAB. With the developed model, the effects of different tilt angles and a different beam angle of luminaire were investigated.

**Keywords** LED light modeling · Simulation · Stadium floodlighting · Energy-efficient

N. S. bin Mat Nanyan (✉) · I. E. Lee · G. C. Chung · D. S. Ong
Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia
e-mail: 1181402127@student.mmu.edu.my

I. E. Lee
e-mail: ielee@mmu.edu.my

G. C. Chung
e-mail: gcchung@mmu.edu.my

D. S. Ong
e-mail: dsong@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_20

## 1 Introduction

Previously, various types of light sources have been used for a stadium floodlighting. However, the performance of lighting to stadium application was seldomly reported. At the early stage in the modern years, a gas discharge lamp has been widely used because of its ability to produce higher light intensity required for a stadium. Reference [1] used HQL type mercury vapor lamp to design a lighting system for a stadium. However, the illumination requirement was based on 250 lx only and is no more practical and far below today's requirements.

As technology changes and improves, metal halide has been widely used as a lighting source for a stadium. The metal halide type is a discharge lamp that uses mercury vapor and metal halide additives at high pressure in quartz or ceramic arc tube contained within a glass or quartz glass outer envelope [2]. It is similar to a mercury vapor lamp but contains additional metal halide compounds in the quartz arc tube, which improve the efficiency and color rendition of the light.

With the invention of LED lighting, the industry has an option to be used as a source of light due to the characteristic of LED lighting that consumes less energy. Reference [3] has reported that LED lighting has smaller energy consumption than metal halide, which proves that LED can be used to save energy usage in stadium applications. However, the methodology is based on comparing the energy consumption between metal halide and LED system only. The author found that uniformity performance is slightly lower. Hence, an approach to improving the uniformity of LED lighting is becoming the most critical challenge in LED applications. The behavior of LED lighting needs to be understood to improve their visual performance.

Assessment of the performance of LED lighting to improve stadium illuminance was done by [4]. The author proposed introducing LED lighting on the existing stadium as a secondary source of light, which already uses metal halide as a stadium pitch floodlighting. The simulation result found that illuminance can be improved with the introduction of LED lighting. However, this is a hybrid system containing two types of lighting, which have slight effects on energy consumption.

The approach to finding appropriate aiming schemes of luminaires using genetic algorithm and mathematical method has been discussed in [5–7]. However, the finding was based on metal halide lamp and not on LED luminaires.

In this study, a computational model of LED luminaire was developed using MATLAB. The effects of different tilt angles and different view angles of luminaire were investigated with the developed model. The model was developed using single LED chip data from the datasheet, which then was assembled as a single luminaire. The evaluation was using from one luminaire and increased gradually until 28 numbers of luminaires.

Stadium lighting specifications have to adapt to the change of broadcast technology. Due to this requirement, the specifications set by the International Federation of Association Football (FIFA) are high. Hence, this research study uses FIFA specification as a guideline to measure the visual performance of LED lighting as in [8].

## 2 Methodology

### *2.1 Modeling LED*

Light is electromagnetic waves that their wavelength is in the visible spectrum. For lighting designers, irradiance is essential. Before irradiance of a light source can be modeled, the light source's basic properties need to be determined. Luminous intensity is one of the fundamental properties of the LED light. It is vital to model the irradiance of light. Luminous intensity is depending on luminous flux produce by a light source. Luminous flux is how much light is perceived by humans that come from a source into a given solid angle. The equation to quantify luminous intensity is given as

$$I(O) = \Phi / \Omega \tag{1}$$

$\Phi$ is luminous flux from a source measured in lumen (lm) and $\Omega$ is a dimensional angle called solid angle measured in steradian (sr). For an LED, it is in hemisphere shape so the solid angle can be calculated using

$$\Omega = 2\pi(1 - \cos\theta) \tag{2}$$

View angle (will be referred to as VA) or also known as beam angle from an LED light source, is described as in Fig. 1 below. Hence, please note that θ is half of the VA for that particular LED.

The equation to simulate the radiation pattern of LED light is given by

$$I(\phi) = I(O)\cos^m(\phi) \tag{3}$$

where $I(0)$ is the center luminous intensity of an LED that can be calculated using Eq. (1) and $m$ is the order of Lambertian radiation emission. $m$ can be calculated by the semi-angle (view angle) at half illuminance of LED $\emptyset_{1/2}$ and is given by

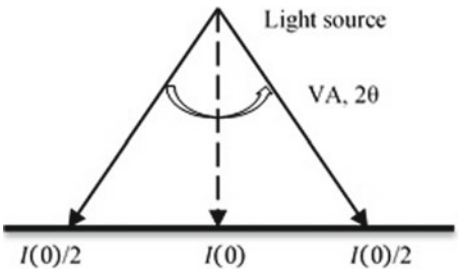



**Fig. 1** Description of VA

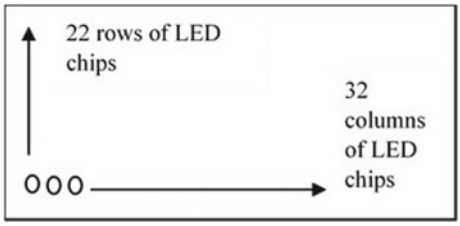



**Fig. 2** Illustration of a complete set of a luminaire

$$m = -\ln 2 / \ln \cos(VA/2) \tag{4}$$

## 2.2 Luminaire

LED chips were then arranged as an array of 32 × 22 LED chips with gap 2 cm in between chips to form a complete set of a luminaire, as illustrated in Fig. 2. The total quantity of LED chips that were used to form a luminaire were 704 pieces.

Three types of luminaire's VA configurations being used were 30°, 60°, and 120°. These types of luminaires were chosen based on the availability in the current market. For the initial evaluation, a luminaire with a single type of VA been used to evaluate its light output. The luminaires were mounted on a high mast located at the center of each quadrant (total four high masts) of a football field. In this paper, three sets of luminaires setup been studied. The three setups were:

- All 28 luminaires (on each quadrant) with VA 30°.
- All 28 luminaires (on each quadrant) with VA 60°.
- All 28 luminaires (on each quadrant) with VA 120°.

All of the luminaires being tilted uniformly for each setup. The tilt angle has been used were 30°, 40°, and 50°. Each of the luminaire types will have a different effect on the light intensity and also light distribution, as simulated in Fig. 3. The simulated pattern compared to the radiation pattern characteristics in the LED chip datasheet as in [9].

## 2.3 Evaluation Area

The layout of the evaluation area is shown in Fig. 4, a standard from FIFA guidelines. The width of the field been used in the simulation is 100 m, and breadth is 64 m. The field was segregated into four quadrants named as 'Quadrant I', Quadrant II', 'Quadrant III' and 'Quadrant IV'. The simulation will be based on a quadrant as the rest of the quadrants were mirrored vertically and horizontally.

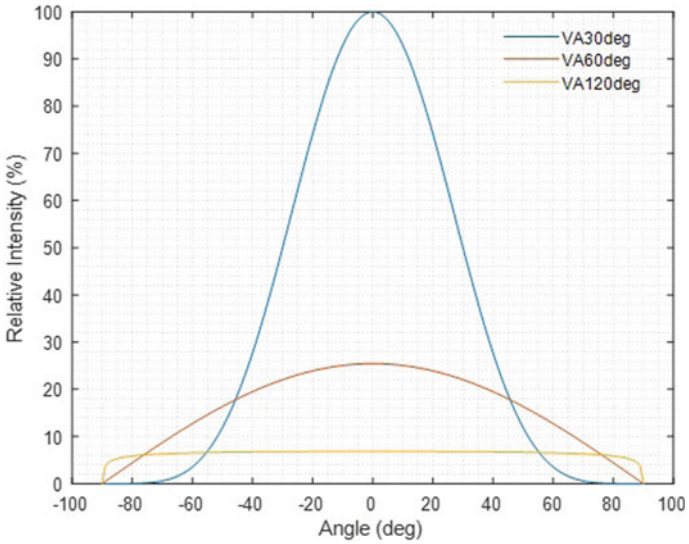



**Fig. 3** Illustration luminaire's intensity with different VA

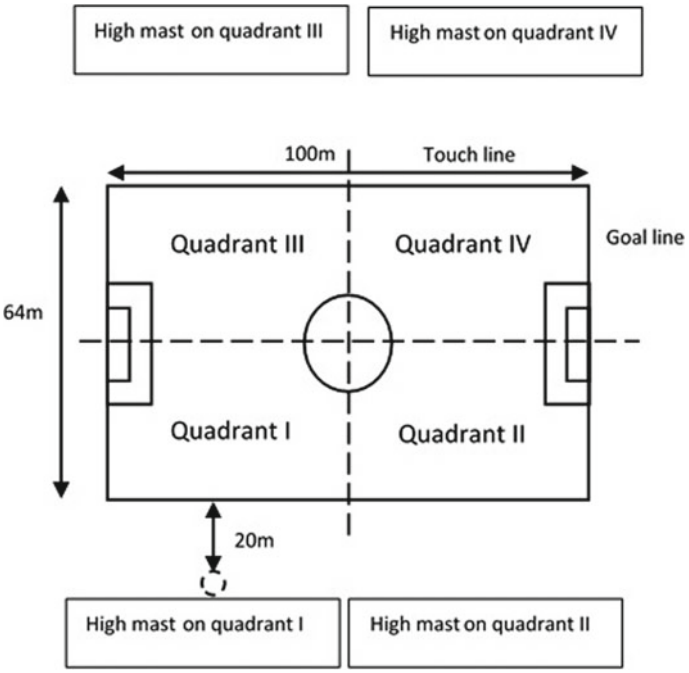



**Fig. 4** The layout of a football field and the position of four high masts

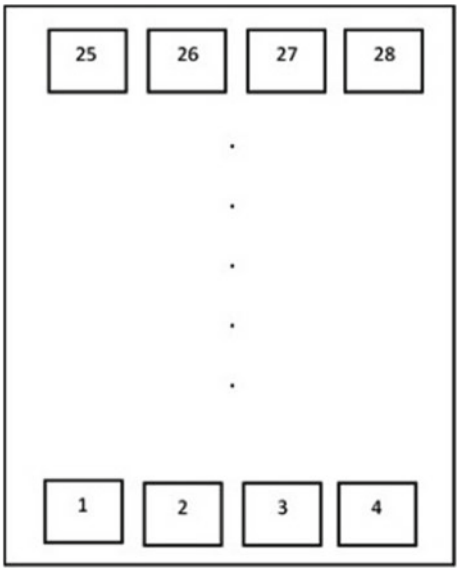



**Fig. 5** Luminaires arrangement on a high mast

There were four high masts positioned at the center at each quarter of the field. They were 20 m apart from the touchline and 45 m height. For the initial evaluation, a total of 28 luminaires were arranged, as in Fig. 5 arrangement. Four luminaires installed in typical arrays on seven horizontal rows. They were increased one by one to simulate their output.

### *2.4 Horizontal Illuminance*

A horizontal illuminance $E_h$ at a point (x, y) is given by

$$E_h = I(O)\cos^m(\phi)/D_d^2\cos(\Psi) \tag{5}$$

Ø is tilt angle of the luminaire, Dd is the distance between luminaire and a surface while Ψ is the angle of incidence as illustrated in Fig. 6.

## 3 Performance Indicator

Several parameters need to be considered and computed during the design process to quantify the light quality and need to be measured and verified after installation.

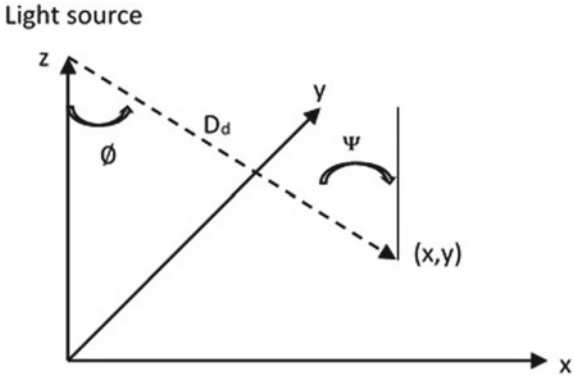



**Fig. 6** Three-dimensional illustration of a geometric involved to calculate the horizontal illuminance on a coordinate (x, y)

The main objective of stadium lighting is to illuminate events to digital video quality for the media without creating nuisance glare for the players and adding spill light or glare to the spectators and the surrounding environment. Table 1 above showing the requirement for stadium lighting as in [8].

Measures of uniformity quantify the variations of illuminance over the playing area. Two types of uniformity stated in the FIFA guideline. First is uniformity $U_1$ that is the ratio between minimum horizontal illuminance and maximum horizontal illuminance. The second is uniformity $U_2$, which is the ratio between minimum

**Table 1** Stadium lighting parameters requirements

| Lighting class | | Vertical Illuminance | | | Horizontal Illuminance | | |
|---|---|---|---|---|---|---|---|
| | | $E_v$ cam avg (lx) | Uniformity | | $E_h$ avg (lx) | Uniformity | |
| | | | U1 | U2 | | U1 | U2 |
| Class V(International) | Fixed Camera | 2,400 | 0.5 | 0.7 | 3,500 | 0.6 | 0.8 |
| | Field camera (at pitch level) | 1,800 | 0.4 | 0.65 | | | |
| Class IV(National) | Fixed Camera | 2,000 | 0.5 | 0.65 | 2,500 | 0.6 | 0.8 |
| | Field camera (at pitch level) | 1,400 | 0.35 | 0.6 | | | |
| Class III (National games) | | NA | | | 750 | NA | 0.7 |
| Class II (League and club) | | NA | | | 500 | NA | 0.6 |
| Class I (Training and recreation) | | NA | | | 200 | NA | 0.5 |

horizontal illuminance and average horizontal illuminance. $U_1$ and $U_2$ are applicable for televised event specifications. While only $U_2$ applicable for non-televised events.

In this study, the parameters and data that are considered and collected were for horizontal illuminance. While the uniformity ($U_1$ and $U_2$) can be computed from illuminance data taken. These parameters were then compared to the FIFA specifications in Table 1.

Uniformity is based on horizontal illuminance at the grid points over the playing area, as described in [8]. Hence uniformity is dependent on any change in the illuminance distribution over the playing area.

## 4 Result and Discussion

### *4.1 Light Output*

The light output data for one luminaire was taken to be analyzed to understand the characteristics of LED lighting. The number of luminaires will be gradually increased until the maximum of 28 luminaires. The light output in a single quadrant and the whole football field consists of all four quadrants being analyzed. With a single quadrant, the high mast placed at the center of a quadrant, the horizontal illuminance been simulated.

From Fig. 7 above, when all the luminaires being tilted at 40°, it can be seen that for all the three setups, the light intensity was concentrated at the center (aiming point) for VA 30°, but illuminance became lower at the side. As the VA became wider, the light intensity started to spread. As a result, it can be seen that the horizontal illuminance started to drop, as in Fig. 7b and c. From all the setups, light radiation was not changed with the increase of luminaires but only increased illuminance values.

With luminaires on the other high mast (all quadrants) been turned on, the illuminance was simulated as well.

When all the luminaires being activated on each of the high masts (all quadrants), the light outputs were as in Fig. 8. The light intensity for luminaires with VA 30° was concentrated at the center and had the highest average illuminance value but sharply dropped toward the area circled as in Fig. 8a. The light intensity for luminaires with VA 60°, were concentrated at the center as well and had high illuminance value but gradually dropped toward the area circled as in Fig. 8b. For luminaires with VA 120°, the average illuminance value was the lowest but had a small margin between the maximum and minimum average horizontal illuminance. Now we know that luminaires with VA 30 and 60° have the advantage of complying with illuminance requirements but have limitations in spreading the lights uniformly.

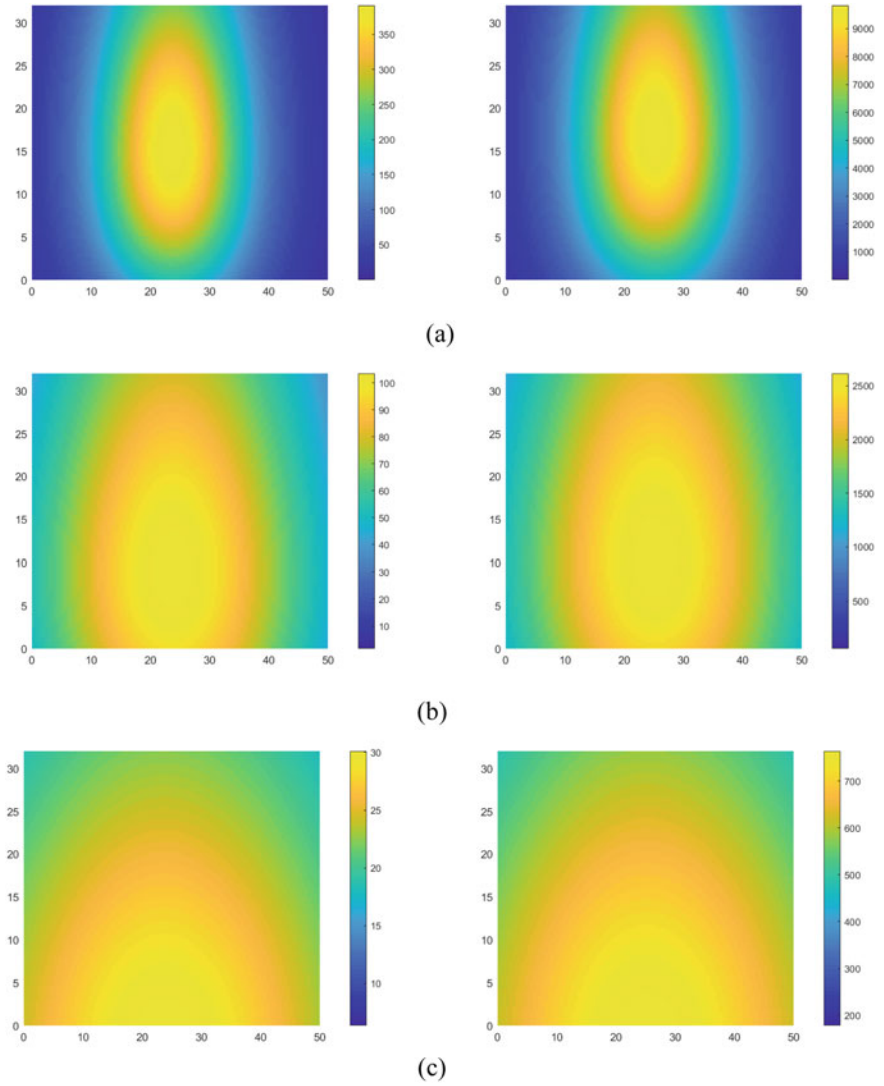



**Fig. 7** The light output on a quadrant for: **a** the first setup, with one and 28 luminaires, **b** the second setup, with one and 28 luminaires, and **c** the third setup, with one and 28 luminaires on one quadrant

## *4.2 Horizontal Illuminance*

Horizontal illuminance is a measure of light reaching a horizontal plane. The effect of tilting and adding the numbers of luminaires to the horizontal illuminance was investigated.

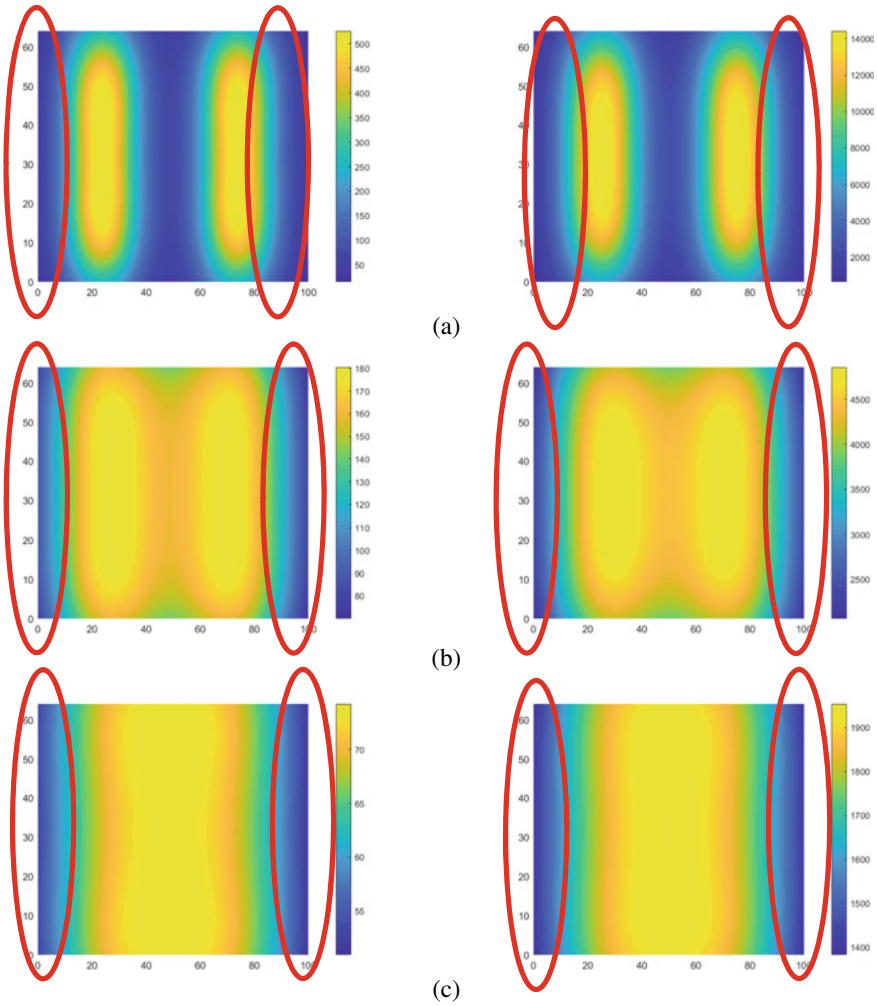



**Fig. 8** Light output from a single luminaire mounted on top of four high masts for: **a** the first setup, with one and 28 luminaires, **b** the second setup, with one and 28 luminaires and **c** the third setup, with one and 28 luminaires on all four quadrants

From Fig. 9, it can be seen that the average horizontal illuminance was increased with the increasing quantity of luminaires. Luminaires with tilt angle 40° and 50° always produce the highest output in terms of illuminance. Luminaire with VA 30° and 60° able to meet the highest FIFA average horizontal illuminance specification (Class V = 3,500 lx), with a minimum quantity of luminaires. However, luminaires with VA 120° did not meet the Class V FIFA specification, although with the maximum quantity of luminaires as in Fig. 9c.

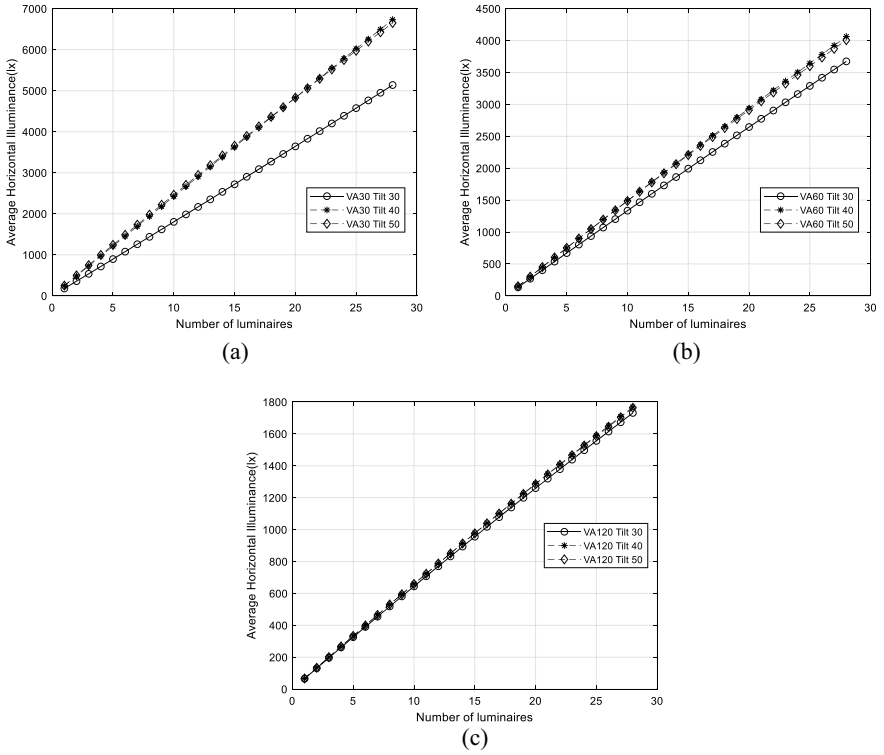



**Fig. 9** The average horizontal illuminance with the increased quantity of luminaires and tilt angle for: **a** the first setup, **b** the second setup, and **c** the third setup

From the results of light output and horizontal illuminance above, it shows that single type luminaires setup can meet the average horizontal illuminance for luminaires with VA 30° and 60°. However, there was a limitation as the light tends to drop at the sides of the aiming point of that particular setup happened even by increasing the number of luminaires as the light will concentrate at the aiming point and jeopardize the overall uniformity requirement. This simulation has demonstrated the approach in assessing lighting performance, and there is a possibility to improve the visual performance.

## *4.3 Uniformity*

From the light output discussed above, it is known that there were limitations that produce low illuminance spots with the current luminaires setup. The effect of tilting and adding the numbers of luminaires to the uniformity was investigated as well.

From Fig. 10a shows that luminaires with VA 30° have the lowest uniformity compared to the other types of luminaires. In contrast, luminaires with VA 120° produced the best uniformity output if compared within those three types of luminaires. Additionally, in terms of tilting the luminaires, 30° tilt of the luminaires always produced relatively the best result compared to the other two tilt angle for all types of luminaires. However, these uniformity results were only intended to

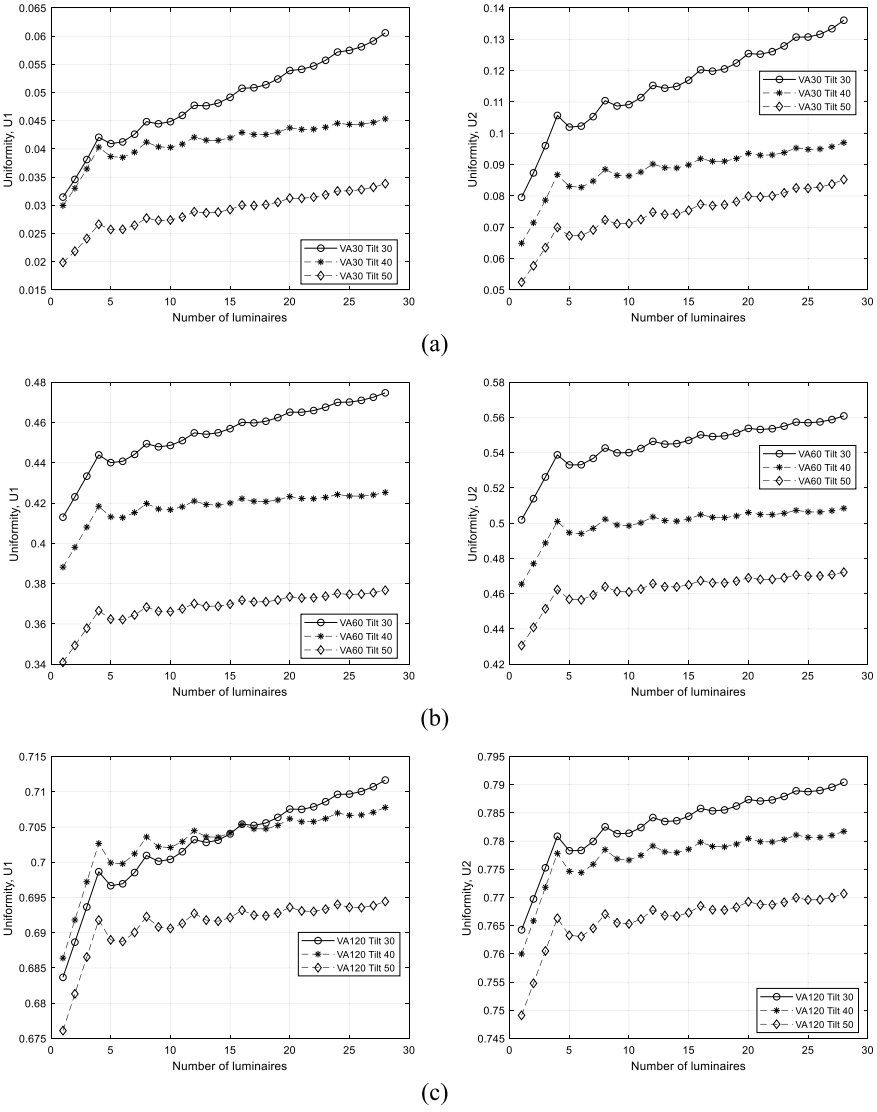



**Fig. 10** The uniformity $U_1$ and $U_2$ with the increasing quantity of luminaires and tilt angle for luminaires with: **a** VA 30°, **b** VA 60° and **c** VA 120°

show the effect of tilting and VA of luminaires to uniformity output. Regarding the FIFA lighting specifications, as in Table 1, none of the luminaires can fulfill uniformity requirements as the uniformity computed was on the overall football field. It is inclusive of the low illuminance spots, as in Fig. 8. Hence, there is an opportunity to improve visual performance by optimizing the architectural design focusing on the lower illuminance spots.

## 5 Conclusion

This paper has discussed the characteristics of LED lighting that are used as stadium lighting. Luminaire with narrow view angle gives higher illuminance value, and the light intensity is concentrating at the point of aiming. In contrast, wider view angle luminaire gives better uniformity as it spread lights well from the point of aiming toward the sides.

It was also found that single VA luminaires were not sufficient enough to satisfy FIFA stadium lighting specifications. Luminaires with VA 30° have the advantage of producing the highest illuminance value. In comparison, luminaires with VA 60° have the advantage of producing better uniformity. For luminaires with VA 120° have better uniformity but cannot comply with horizontal illuminance requirements with the minimum numbers of luminaires. Hence, there is an opportunity further to investigate both types of luminaires VA 30° and 60° and evaluate the effects of combining the different VA luminaires in order to meet the horizontal illuminance and uniformity requirements.

## References

1. Unver MU, Imal N (1999) Lighting design for a football field. In: Eleco'99 international conference on electrical and electronics engineering
2. The institution of lighting engineers (2005) The outdoor lighting guide, 1st edn. Taylor & Francis, UK
3. Polycarpou A, Yiannou I, Christofides N (2016) Comparison of MH and LED performance for sport lighting application. In: Mediterranean conference on power generation, transmission, distribution and energy conversion (MedPower), pp 21–25
4. Mangkuto RA, Rachman AP, Aulia AG, Asri AD, Rohmah M (2018) Assessment of pitch floodlighting and glare condition in the main stadium of Gelora Bung Karno, Indonesia Meas J Int Meas Confed 117(July 2016):186–199
5. Petranovi D (2015) Football stadium floodlight aiming by using a genetic algorithm with multi-step approach. Polytech Des 3(2):135–143
6. Xiao H, Fang J, Zhu P, Kang Q (2014) Application of genetic algorithms in football field lighting for energy-saving. In: 26th Chinese Control and Decision Conference (CCDC), pp 664–669
7. Nath D, Mazumdar S, Chandra J, Bag A (2016) A rough set-based method for aiming angle tuning of luminaires for outdoor sports lighting. Light Res Technol 48:126–154

8. FIFA (2007) Football stadiums technical recommendations and requirements, 4th edn
9. Package CS, Tx L "LUXEON V2." https://www.lumileds.com/wp-content/uploads/files/DS177-luxeon-v2-datasheet.pdf. Accessed on 6 July 2020

# Split Balancing (sBal)—A Data Preprocessing Sampling Technique for Ensemble Methods for Binary Classification in Imbalanced Datasets


**Chongomweru Halimu and Asem Kasem**

**Abstract** The problem of class imbalance in machine learning occurs when there is a relatively big disproportional distribution of classes in the data for classification tasks. In many real-world domains, such as in healthcare, finance, and predictive maintenance, the number of data points of a less important class (usually the negative class) is much higher than the class of greater interest (usually the positive or target class). This affects the ability of many learning algorithms to find good classification models. To address that, many approaches for solving this problem have been proposed, prominently including ensemble methods integrated with sampling-based techniques. However, these methods are still prone to the negative effects of sampling-based techniques that alter class distributions via over-sampling or under-sampling, which can lead to overfitting or discarding useful data, respectively, and thus affect performance. In this paper, we propose a new data preprocessing sampling technique dubbed as (sBal) for ensemble methods for binary classification in the case of imbalanced datasets. Our proposed method first turns the imbalanced dataset into several balanced bins/bags. Then multiple base learners are induced on the balanced bags and finally, the classification results are combined using a specific ensemble rule. We evaluated the performance of our proposed method on 50 imbalanced real-world binary datasets and compared its performance with well-known ensemble methods that utilize data preprocessing techniques namely SMOTE-Bagging, SMOTEBoost, RUSBoost, and RAMOBoost. The results reveal that the proposed method brings considerable improvement in classification performance relevant to the compared methods. We performed statistical significance analysis using Friedman's non-parametric statistical test with Bergman post-hoc test. The analysis showed that our method performed significantly better than the majority of the methods across many datasets, suggesting a better preprocessing approach than the ones used in compared methods. We also highlight possible extensions to the method that can improve its effectiveness.

C. Halimu (✉) · A. Kasem
Universiti Teknologi Brunei, Jalan Tungku Link BE1410, Gadong, Brunei Darussalam
e-mail: p20181008@student.utb.edu.bn

A. Kasem
e-mail: asem.kasem@utb.edu.bn



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_21

**Keywords** Data sampling · Ensemble methods · Imbalanced datasets · Split balancing · RAMOBoost · SMOTEBagging · SMOTEBoost

## 1 Introduction

Class imbalance is a common challenge in the machine learning community and academia. It occurs in a dataset where the class of interest (positive or minority class) has fewer instances than the other class (negative or majority class) [1]. In real-world applications, the problem of class imbalance has affected many different domains such as; credit card fraud detection, medical diagnosis, anomaly detection, text classification among others. This is because the class of interest is usually rare as compared to the majority class. For example, in some countries, the number of male drivers is much more than female drivers, yet they are equally important in assessing the causes of road accidents. To address the problem of class imbalance, various methods have been proposed, and these may be categorized into data level, algorithm level, and ensemble level methods [2].

The data level methods try to rebalance the class distribution through sampling, which includes both under-sampling and oversampling based techniques. For the under-sampling-based techniques like Random Under-Sampling (RUS), they alter the original class distribution of the dataset by randomly eliminating some instances from the majority class. Whereas, for the over sampling-based techniques, they generate new instances to the minority class to balance the class distribution of the dataset. Different variations of over-sampling-based techniques have been proposed and evaluated in the literature [3, 4]. The most common ones include Random Over Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE) among others. Furthermore, the algorithm level methods focus on manipulating the algorithm through cost-sensitive learning [5], by raising the cost of misclassifying minority instances and reducing that of misclassifying the majority instances to bias the algorithm towards the positive class which is always the class of interest. Examples of the most commonly used algorithm level method include AdaBoost. Finally, the ensemble level methods, generate various classification models using a set of base learners and then use a given combination rule to combine their decisions into a single classification result. When compared with the data level and algorithm level, ensemble-based methods such as bagging, and boosting are considered the most successful and commonly used in practical applications [6].

However, most of the above-mentioned approaches might still face unanticipated challenges when employed to solve the problem of class imbalance. For instance, the data level oversampling-based methods might increase the likelihood of overfitting due to the generation of new instances to the minority class. Whereas, the under sampling-based methods might discard some potentially valuable data that might be very important during the training process. Furthermore, ensemble-based methods such as bagging and boosting, are not always effective at handling class imbalance problems unless combined with data level sampling techniques [7]. For example,

SMOTE is combined with Bagging to form SMOTEBagging, with Adaboost to form SMOTEBoost, UnderSampling with Bagging forms UnderBagging. Despite using data sampling-based techniques as a precursor of the ensemble methods in handling class imbalance problem, they are still prone to the shortfalls of sampling-based methods such as overfitting and loss of potentially useful data, since they still use the sampling techniques to generate balanced bootstraps in each iteration of the bootstrap aggregating process. Hence, a wider study is still needed for handling data preprocessing sampling problems for ensemble methods in the binary classification of imbalanced datasets.

In this paper, we propose a new Split Balancing data preprocessing sampling technique dubbed as (sBal) for ensemble methods in the binary classification of imbalanced datasets. Our proposed method first splits the majority instances of the imbalanced dataset into multiple splits/bags and then balances the class distribution of each split by joining it with all the instances of the minority class to form a series of balanced bags. Then multiple base learners are induced on the balanced bags to generated multiple models which are then combined using a specific ensemble rule.

We conducted comprehensive experiments on 50 real-world imbalanced binary datasets obtained from KEEL [8], and UCI online repository and compared our proposed methods with well-known ensemble methods that utilize data sampling preprocessing techniques which include; SMOTEBagging, SMOTEBoost, RUSBoost, and RAMOBoost. The findings show that our proposed methods method performed significantly better than most of the existing methods. To validate our results, we went ahead to carry out a non-parametric statistical test, to ascertain if there exists any significant difference in performance between our proposed method and the existing methods. We observed significant differences in performance between sBal and the majority of the methods being studied.

In summary, the reminder part of this paper is organized as follows: In Sect. 2 we present the related work on data sampling techniques and ensemble methods that utilize data sampling as a precursor for solving the problem of class imbalance. Section 3 presents our proposed methods. The detailed experimental design is presented in Sect. 4. We then present results analysis and discussion in Sect. 5. Finally, we make our conclusions and propose future works in Sect. 6.

## 2 Related Work

The problem of class imbalance has caught much attention of researchers in the industry and academia. As a result, numerous solutions for tackling this problem have been proposed and they may fall into different categories, which include; the data level, algorithm, and ensemble level methods. In this paper, our focus is on data level-based methods more especially sampling techniques with ensemble methods. In this section, we present a summary of existing ensemble and sampling-based methods for handling the class imbalance problems that are much related to our study.

Data sampling and ensemble learning are among the most commonly used methods in solving the problem of class imbalance [9]. The sampling methods aim at balancing the class distribution of the imbalanced dataset and these can be divided into two types; Oversampling based methods, which introduce new instances into the minority class and Undersampling, which discard instances from the majority class to obtain a balanced class distribution. Several studies in the literature [10–12], have proposed different variations of over sampling-based methods. The simplest oversampling method is the Random Over Sampling (ROS), it duplicates instances from the minority class to balance the class distribution of the imbalanced dataset. Other studies [13, 14], have similarly prosed different under-sampling based approaches such as Random Under Sampling (RUS) for solving the problem of class imbalance. Japkowicz in [15], discussed both the under-sampling and sampling strategies and pointed out the fact that both strategies were effective and further noted that, using sophisticated data sampling techniques may not give any clear advantage in solving the problem of class imbalance. In another literature, Dittman et al. [16], carried out a comparative study of different sampling methods (RUS, ROS, and SMOTE) used in the classification of class imbalanced problems in the field of Bioinformatics. Based on their statistical analysis and findings, they reported that RUS is the most preferred technique this is because of its ability to reduce the dataset size and the subsequent computational overhead. The main shortfalls of the sampling-based methods are that they alter the original class distribution of the datasets which can lead to unforeseen errors. For instance, oversampling may lead to overfitting whereas the under-sampling methods may discard potentially useful data.

On the other hand, ensemble methods work towards improving classification performance and generalization ability on the classification of future instances [6]. The majority of these ensemble methods are combined with data preprocessing sampling techniques such as under sampling and over sampling-based techniques [2, 12]. Bagging and Boosting are some of the methods utilizing data sampling techniques and they have been widely used in the classification of class imbalance problems [9]. Researchers in [17], carried out an empirical study, comparing different sampling methods with boosting for enhancing the performance of decision trees in the identification of defective modules in the software. In their results, they indicated that sampling techniques helped in improving the performance of such models. Chawla et al. in their study [18], proposed a new SMOTEBoost ensemble method for classification of class imbalance problems that is based on the SMOTE sampling technique and boosting algorithms. In another empirical study [11], a Random Under Sampling Boost (RUSBoost) ensemble is proposed which is a result of combining the RUS technique with the boosting method. Their findings show a comparable performance between RUSBoost and SMOTEBoost.

However, most of these ensemble-based methods, are based on data sampling techniques and therefore they may alter the class distribution of the original imbalanced datasets by either discarding some of the instances from the majority class (under-sampling) or by generating more instances for the minority class (oversampling) which might cause overfitting since most learning algorithms tend to pay much

attention to the replicated minority instances [3]. Furthermore, for the case of traditional bagging and boosting ensemble-based methods, they might still suffer from the challenge of class imbalance. This is because for each of the iteration (in both boosting and bagging based methods) the class distribution in each sampled subset in a certain iteration is the same as that of the original dataset.

It is, therefore, prudent to have a sampling method that can easily overcome the shortfalls of the previously studied methods, by balancing the class distributions of the imbalanced dataset without creating new data that might lead to overfitting or without discarding data that might be potentially useful during the learning process.

Our proposed Split Balancing data preprocessing sampling technique (sBal), address the previous shortfalls by first splitting the majority instances of the imbalanced dataset into multiple splits/bags and then balances the class distribution of each bag by joining it with all the instances of the minority class to form a series of balanced bags as depicted in Fig. 1. Then multiple base learners are induced on the balanced bags to generate multiple models which are then combined using a given ensemble combination rule to get the final classification results.

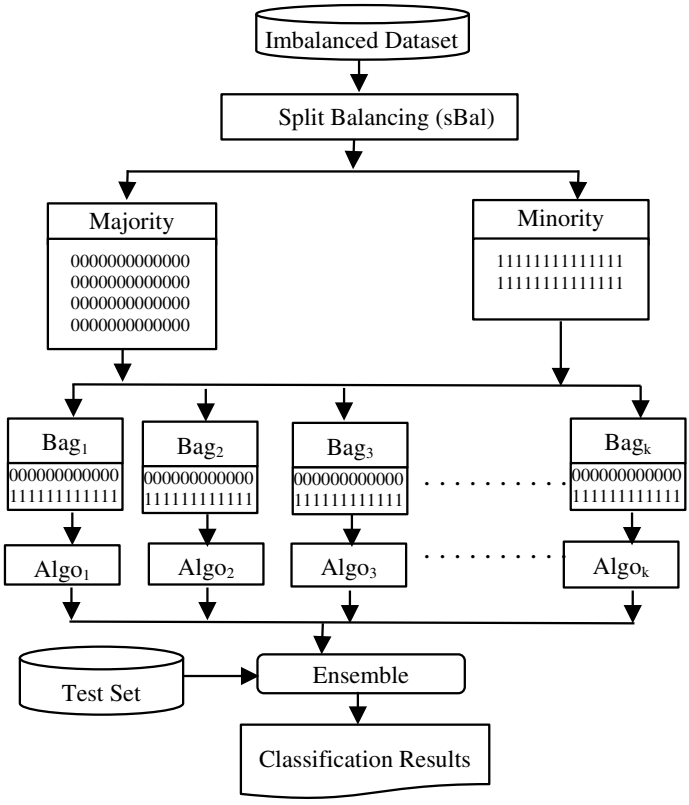



**Fig. 1** Proposed sBal method

Our proposed methods greatly differ from the previously existing methods in many different ways, first, we don't discard any potentially useful data as it has always been in most under sampling-based methods such as RUS, similarly, we minimize the chances of overfitting since we do not replicate or introduce new data like in most over sampling-based techniques such as ROS. Additionally, it has minimal computational overhead as compared to SMOTE that must first calculate the k distance before creating a new instance.

## 3 Proposed Method

Traditional machine learning methods for binary classification are designed to perform better when presented with balanced datasets [19]. However, the majority of them tend to perform poorly when a dataset is imbalanced since the cost of classifying the positive (minority) class is always much higher as compared to the cost of classifying the negative (majority) class [20]. As a result, various techniques that try to balance the class distribution of an imbalanced dataset such as over-sampling and under-sampling-based techniques have been proposed. Furthermore, several ensemble methods have been combined with different sampling-based techniques to solve the problem of class imbalance. However, they might still inherit some weaknesses of sampling-based methods as earlier discussed.

Our proposed method (Fig. 1), tries to address the possible shortfalls of the traditional methods for handling the problem of class imbalance by first turning the class imbalance problem into a series of multiple balanced problems without creating/introducing any new instances into the minority class or discarding any original instances from the majority class. Considering an imbalanced dataset where the number of instances in the majority (negative) class is more than the instances in the minority (positive) class, we randomly split the instances of majority class into multiple $k$ bins/bags, where the number of $k$ is determined by the ratio of the majority class instances to the minority class instances. Each bag will contain different instances from the majority class where the number of majority instances in a given bag is determined by the size of instances in the minority class. We then join a copy of all instances of the minority class to each of the $k$ bags to form a series of balanced $k$ bags as shown in Fig. 1. Finally, we apply classification on each of the balanced $k$ bags using multiple base learners, the obtained classification results from the multiple classifiers are combined using a recommended ensemble combination rule (majority voting) to form a new ensemble method that is based on sBal as a data preprocessing sampling technique.

## 4 Experimental Design

In this section, we present details of various experiments evaluating the effectiveness of our proposed technique (sBal) and compare its performance against popular existing ensemble methods that utilize data level preprocessing sampling techniques in handling class imbalance problem which include SMOTEBagging, SMOTEBoost, RUSBoost, RAMOBoost, and traditional Bagging and RF methods on 50 imbalanced multi-domain real-world binary datasets (Table 1) in terms of AUC.

### *4.1 Datasets*

Table 1 summarizes all the datasets used in this study along with their respective properties. For each of the datasets, we show the dataset ID (ID), Dataset name (Dataset), Total number of instances (#Inst), Percentage of Majority instances (%Maj), Percentage of minority instances (%Min) and Imbalance Ratio (IR). In this study, we consider IR as the ratio of percentages of Majority to Minority instances, i.e. $IR = \frac{\%Maj}{\%Min}$. All the datasets are real-world coming from different domains, they were prepared for binary classification tasks in mind and they are publicly available on KEEL [8], and UCI repositories.

### *4.2 Evaluation Metrics*

Accuracy is a famous evaluation metric and commonly used for the evaluation of several machine learning problems. However, studies have reported the ineffectiveness of accuracy in the evaluation of class imbalance problems [21]. As a result, researchers and practitioners have opted for alternative evaluation metrics such as G-Means, F-Measure, Recall, Precision, Area Under Curve (AUC), and Mathew Correlation Coefficient (MCC) for evaluation of class imbalance problems. The majority of researchers in the ML community are preferring AUC [21] and MCC [22] over other metrics. In a recent study [23], researchers empirically compared and evaluated the use of AUC and MCC metrics in the evaluation of class imbalance problems. In their findings, they established that AUC was statistically consistent and more discriminating than MCC; hence suggesting that AUC is a better measure than MCC to be used for evaluating binary classification with imbalanced datasets. In that regard, we chose to use AUC as an evaluation metric for this study.

**Table 1** Summary of real-world binary imbalanced datasets sorted by IR

| ID | Datasets | #Feat | #Inst | %Maj | %Min | IR | ID | Datasets | #Feat | #Inst | %Maj | %Min | IR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | monks_prob_2 | 15 | 432 | 67.1 | 32.9 | 2.0 | D26 | cardiotocography_c1c3 | 21 | 1831 | 90.4 | 9.6 | 9.4 |
| D2 | vertebral_column | 6 | 310 | 67.7 | 32.3 | 2.1 | D27 | cardiotocography_c2c3 | 21 | 1831 | 90.4 | 9.6 | 9.4 |
| D3 | credit-g | 19 | 1000 | 70.0 | 30.0 | 2.3 | D28 | vowel0 | 13 | 988 | 90.9 | 9.1 | 10.0 |
| D4 | car_evaluation | 6 | 1728 | 70.0 | 30.0 | 2.3 | D29 | glass-0-1-6_vs_2 | 9 | 192 | 91.2 | 8.9 | 10.3 |
| D5 | breast_cancer | 13 | 286 | 70.3 | 29.7 | 2.4 | D30 | ecoli-0-1-4-7_vs_2-3-5-6 | 7 | 336 | 91.4 | 8.6 | 10.6 |
| D6 | indian_liver_patients | 10 | 583 | 71.4 | 28.6 | 2.5 | D31 | climate_model | 18 | 540 | 91.5 | 8.5 | 10.7 |
| D7 | haberman | 3 | 306 | 73.5 | 26.5 | 2.8 | D32 | led7digit-0-2-4-5-6-7-8-9_v | 7 | 443 | 91.7 | 8.4 | 11.0 |
| D8 | page-blocks-1-3_vs_4 | 10 | 472 | 94.1 | 28.0 | 3.4 | D33 | glass-0-6_vs_5 | 9 | 108 | 91.7 | 8.3 | 11.0 |
| D9 | hepatitis | 19 | 155 | 79.4 | 20.7 | 3.8 | D34 | glass-0-1-4-6_vs_2 | 9 | 205 | 91.7 | 8.3 | 11.1 |
| D10 | spect_heart | 22 | 267 | 79.4 | 20.6 | 3.9 | D35 | glass2 | 9 | 214 | 92.1 | 7.9 | 11.6 |
| D11 | cardiotocography_c1c2 | 21 | 1950 | 84.9 | 15.1 | 5.6 | D36 | cleveland-0_vs_4 | 13 | 173 | 92.5 | 7.5 | 12.3 |
| D12 | balance_scale_BL | 4 | 337 | 85.5 | 14.5 | 5.9 | D37 | ecoli-0-1-4-6_vs_5 | 6 | 280 | 92.9 | 7.1 | 13.0 |
| D13 | balance_scale_BR | 4 | 337 | 85.5 | 14.5 | 5.9 | D38 | shuttle-c0-vs-c4 | 9 | 1829 | 93.3 | 6.7 | 13.9 |
| D14 | internet_ad_cfs | 24 | 3279 | 86.0 | 14.0 | 6.1 | D39 | seismic-bumps | 18 | 2584 | 93.4 | 6.6 | 14.2 |
| D15 | ecoli-0-3-4_vs_5 | 7 | 200 | 90.0 | 10.0 | 9.0 | D40 | yeast-1_vs_7 | 7 | 459 | 93.5 | 6.5 | 14.3 |
| D16 | yeast-2_vs_4 | 8 | 514 | 90.1 | 9.9 | 9.1 | D41 | cervical_cancer_risk_facto- | 33 | 858 | 93.6 | 6.4 | 14.6 |
| D17 | ecoli-0-6-7_vs_3-5 | 7 | 200 | 90.1 | 9.9 | 9.1 | D42 | glass4 | 9 | 214 | 93.9 | 6.1 | 15.5 |
| D18 | glass-0-1-5_vs_2 | 9 | 172 | 90.1 | 9.9 | 9.1 | D43 | ecoli4 | 7 | 336 | 94.1 | 6.0 | 15.8 |
| D19 | yeast-0-3-5-9_vs_7-8 | 8 | 506 | 90.1 | 9.9 | 9.1 | D44 | glass-0-1-6_vs_5 | 9 | 184 | 95.1 | 4.9 | 19.4 |
| D20 | yeast-0-2-5-7-9_vs_3-6- | 8 | 1004 | 90.1 | 9.9 | 9.1 | D45 | yeast-1-4-5-8_vs_7 | 8 | 693 | 95.7 | 4.3 | 22.1 |
| D21 | ecoli-0-4-6_vs_5 | 6 | 203 | 90.2 | 9.9 | 9.2 | D46 | yeast4 | 8 | 1484 | 96.6 | 3.4 | 28.1 |

(continued)

**Table 1** (continued)

| ID | Datasets | #Feat | #Inst | %Maj | %Min | IR | ID | Datasets | #Feat | #Inst | %Maj | %Min | IR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D22 | ecoli-0-1_vs_2-3-5 | 7 | 244 | 90.2 | 9.8 | 9.2 | D47 | yeast-1-2-8-9_vs_7 | 8 | 947 | 96.8 | 3.2 | 30.5 |
| D23 | glass-0-4_vs_5 | 9 | 92 | 90.2 | 9.8 | 9.2 | D48 | yeast5 | 8 | 1484 | 97.0 | 3.0 | 32.8 |
| D24 | ecoli-0-3-4-6_vs_5 | 7 | 205 | 90.2 | 9.8 | 9.2 | D49 | ecoli-0-1-3-7_vs_2-6 | 7 | 274 | 97.5 | 2.5 | 39.2 |
| D25 | yeast-0-5-6-7-9_vs_4 | 8 | 528 | 90.3 | 9.7 | 9.4 | D50 | yeast6 | 8 | 1484 | 97.6 | 2.4 | 41.4 |

### *4.3 Experiments Design*

We organized our experiments and their analysis into two phases. In the first phase, we carried out experiments on a series of real-world multi-domain binary imbalanced datasets. To assess performance, we used five repetitions of fivefold cross-validation. To ensure fairness and uniformity across all the studied methods, we used Decision Tree (DT) with its default parameters as the base algorithm. We also constructed all the ensembles with an ensemble size of 10 (n_estimators), this is because our method is limited by the number of base estimators it can use since the number is determined by the ratio of majority to minority instances (*k* bags). The size of an ensemble is another subject of discussion, there is limited literature that gives clear direction about how many n estimators should be used in building an ensemble. Thais et al. in their work [24], analyzed the performance of RF ensemble as the number of trees grow from 2 to 4096 across 29 datasets. Their findings indicate that many trees only increased computational cost, and there was no significant performance gain. They also statistically observed that there was no significant difference in performance between using a given number of trees or double.

In the second phase of the experiments, we carried out nonparametric statistical tests to statistically compare our proposed method against the existing methods. The main objective of this analysis is to validate whether there exists any significant difference in performance between the proposed method and the existing methods being studied in terms of AUC.

All experiments were carried out in Python 3.7 using Scikit-learn version 0.21.3, and imbalanced-learn package version 0.5. For statistical significance tests, we used KEEL's non-parametric statistical analysis tool [8], version 3.0.

## 5 Results and Discussions

In this section, we present and discuss the outcomes of our experiments resulting from the fivefold cross-validation runs. We compared the proposed method against traditional Bagging (Bag'g), Random Forest (RF) and well-known ensemble methods that utilize data preprocessing sampling techniques to overcome the challenge of class imbalance, and these include; RUSBoost (RUB), SMOTEBagging (SMTBag), SMOTEBoost (SMTBst), and RAMOBoost (RAMOB). They were evaluated on 50 publicly available binary datasets (Table 1) obtained from KEEL and UCI repository.

Experimental results for all the methods in the study are presented in Table 2, showing the mean AUC values of 5 repeated trials of the 50 imbalanced binary datasets. The results highlighted in bold indicate a higher AUC score of one method as compared to the other methods in the same row for a given dataset. In the case of a tie between methods, it is counted as a win in its respective capacity. The total number of wins (bold) for each method is shown in the last row of the table.

**Table 2** AUC score for sBal against other ensemble methods on 50 real-world binary imbalanced datasets

| ID | Bag'g | RF | RUB | SMTBag | SMTEBst | RAMOB | sBal | ID | Bag'g | RF | RUB | SMTBag | SMTEBst | RAMOB | sBal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 0.508 | 0.373 | 0.388 | 0.427 | **0.862** | 0.859 | 0.462 | D26 | **0.987** | 0.916 | 0.961 | 0.986 | 0.906 | 0.961 | 0.982 |
| D2 | 0.872 | **0.893** | 0.822 | 0.879 | 0.840 | 0.877 | 0.873 | D27 | 0.943 | 0.952 | 0.909 | 0.919 | **0.996** | 0.996 | 0.953 |
| D3 | 0.750 | 0.740 | 0.705 | 0.753 | 0.556 | 0.727 | **0.756** | D28 | 0.975 | 0.978 | 0.949 | 0.939 | 0.942 | **0.996** | 0.973 |
| D4 | 0.887 | 0.884 | 0.844 | 0.893 | 0.964 | **0.967** | 0.900 | D29 | 0.590 | **0.737** | 0.593 | 0.734 | 0.597 | 0.472 | 0.735 |
| D5 | 0.978 | 0.987 | 0.983 | 0.982 | 0.932 | **0.990** | 0.984 | D30 | 0.921 | **0.961** | 0.843 | 0.912 | 0.689 | 0.791 | 0.925 |
| D6 | 0.681 | 0.685 | 0.658 | 0.688 | 0.642 | **0.756** | 0.690 | D31 | 0.820 | 0.866 | 0.857 | 0.887 | 0.697 | 0.815 | **0.919** |
| D7 | 0.621 | 0.653 | 0.623 | 0.613 | 0.597 | **0.682** | 0.662 | D32 | 0.844 | 0.821 | 0.914 | 0.873 | 0.805 | **0.940** | 0.874 |
| D8 | 0.979 | 0.997 | 0.954 | 0.995 | 0.994 | **1.000** | 0.996 | D33 | 0.961 | 0.987 | 0.895 | 0.956 | 0.975 | **1.000** | 0.980 |
| D9 | 0.848 | 0.847 | 0.767 | 0.831 | 0.773 | 0.673 | **0.859** | D34 | 0.616 | 0.691 | 0.675 | 0.698 | 0.447 | 0.763 | **0.810** |
| D10 | 0.763 | 0.762 | 0.748 | 0.768 | 0.660 | **0.841** | 0.793 | D35 | **0.735** | 0.704 | 0.691 | 0.708 | 0.592 | 0.592 | 0.639 |
| D11 | 0.928 | 0.992 | 0.938 | 0.922 | 0.941 | **1.000** | 0.952 | D36 | 0.880 | 0.911 | 0.827 | **0.923** | 0.667 | 0.651 | 0.883 |
| D12 | 0.561 | 0.616 | **0.623** | 0.543 | 0.541 | 0.601 | 0.594 | D37 | 0.872 | 0.893 | 0.822 | 0.881 | 0.901 | **1.000** | 0.908 |
| D13 | 0.464 | 0.481 | **0.694** | 0.553 | 0.540 | 0.512 | 0.632 | D38 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| D14 | 0.934 | 0.934 | 0.922 | 0.943 | **0.974** | 0.966 | 0.951 | D39 | 0.590 | 0.599 | 0.598 | 0.603 | 0.626 | **0.722** | 0.639 |
| D15 | 0.909 | 0.932 | 0.964 | **0.968** | 0.848 | 0.807 | 0.957 | D40 | 0.691 | 0.682 | 0.667 | 0.719 | 0.559 | 0.663 | **0.773** |
| D16 | 0.975 | 0.968 | 0.972 | 0.983 | 0.778 | **0.984** | 0.975 | D41 | 0.902 | 0.893 | 0.918 | 0.884 | 0.861 | 0.800 | **0.926** |
| D17 | 0.890 | **0.939** | 0.877 | 0.866 | 0.875 | 0.835 | 0.890 | D42 | 0.990 | 0.842 | 0.850 | 0.861 | **1.000** | 0.754 | 0.863 |
| D18 | 0.723 | 0.673 | 0.709 | 0.738 | 0.604 | 0.693 | **0.853** | D43 | 0.914 | 0.942 | **0.980** | 0.912 | 0.867 | 0.840 | 0.954 |
| D19 | 0.671 | 0.688 | 0.671 | 0.656 | 0.657 | 0.637 | **0.689** | D44 | 0.939 | 0.896 | 0.951 | 0.941 | **0.986** | **0.986** | 0.951 |
| D20 | 0.899 | 0.895 | 0.893 | 0.909 | 0.784 | 0.869 | **0.922** | D45 | 0.767 | **0.804** | 0.758 | 0.752 | 0.531 | 0.672 | 0.797 |
| D21 | 0.930 | **0.984** | 0.934 | 0.965 | 0.848 | 0.855 | 0.941 | D46 | 0.839 | 0.817 | 0.842 | 0.855 | 0.481 | 0.816 | **0.925** |

(continued)

**Table 2** (continued)

| ID | Bag'g | RF | RUB | SMTBag | SMTEBst | RAMOB | sBal | ID | Bag'g | RF | RUB | SMTBag | SMTEBst | RAMOB | sBal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D22 | 0.894 | 0.891 | 0.959 | 0.850 | 0.491 | 0.929 | **0.996** | D47 | 0.645 | 0.590 | 0.599 | 0.640 | 0.489 | **0.690** | 0.619 |
| D23 | 0.994 | **1.000** | 0.994 | 0.994 | 0.971 | 0.971 | 0.988 | D48 | 0.969 | 0.953 | 0.981 | 0.959 | 0.885 | **0.997** | 0.990 |
| D24 | 0.910 | **0.983** | 0.949 | 0.914 | 0.861 | 0.753 | 0.925 | D49 | 0.889 | 0.958 | 0.945 | 0.933 | 0.865 | 0.913 | **0.978** |
| D25 | 0.810 | 0.839 | 0.793 | **0.850** | 0.548 | 0.774 | 0.837 | D50 | 0.845 | 0.855 | 0.874 | 0.821 | 0.847 | 0.845 | **0.903** |
| | | | | | | | | | 3 | 9 | 4 | 4 | 6 | **17** | 14 |

From Table 2, we observe a better performance of the proposed method (sBal) as compared to all the other methods except for RAMOBoost (RAMOB) where the performance was comparable. RAMOBoost performed better in 17 datasets out of 50, followed by sBal with 14 wins out of 50, RF forest came third with 9 wins, then SMOTEBoost with only 6 wins out of the 50 datasets. RUB and SMOTEBagging had a tie of 4 wins each out of the 50 datasets. Traditional Bagging (Bag'g) registered the least performance with only 3 wins out of the 50 datasets. We expected poor performance from the traditional bagging method this is because it doesn't pay much attention to the class imbalance since it does not put any effort into trying to balance the class distribution of the bags during the bootstrap aggregating process.

Even though RAMOBoost achieved comparable performance to sBal, the competition between the two methods was not fair, because RAMOBoost is considered a robust method, it adaptively ranks instances from the minority class at each training iteration according to the sampling probability distribution which is based on the primary data distribution [25]. However, its idea of ranking instances can be studied further in the future to lay the ground for new methods that are based on instance hardness.

The experimental results indicate that it is reasonable to turn an imbalanced problem into a series of balanced problems and do classification on balanced problems to achieve better results other than introducing new data into the datasets that can cause overfitting as it is for the majority of oversampling based method or without discarding data from the majority class that might be useful during the training process.

## *5.1 Statistical Results and Analysis*

To draw more reliable conclusions, we went ahead to perform statistical tests and analyses to determine if there exists any statistically significant difference in the classification performance as measured by AUC between our proposed method and the other ensemble methods being studied. We utilized the standard methodology proposed in the literature by Demšar [26] for testing statistical significance among multiple methods across many datasets. We carried out a non-parametric (distribution-free) statistical test on all mean AUC results. We chose to use Friedman's test [27] because we do not know the distribution of the values in our analyzed data. For our study, we used Friedman's N × N statistical test to first determine if there exists any statistically significant difference among methods being studied across all the datasets, followed by a post-hoc test that is used to identify specific pairs of algorithms that produce differences. We used the Bergman procedure [28] to compute a probability value (*p*-value) for the test on each pair of methods. Bergman is a post-hoc procedure with high statistical power devoted to multiple comparisons. It thoroughly finds all the possible sets of hypotheses for given comparisons and all those elementary hypotheses that cannot be rejected. All the statistical tests we carried out using KEEL's non-parametric statistical analysis tool [8].

Statistical results are presented in Table 3. We highlight with bold all pairs that comprise the sBal method. In the results column, we indicate whether the null hypothesis ($H_0$) is rejected or accepted, and we are much interested in rejecting the $H_0$ hypothesis which states that all the methods are comparable and the observed differences in their ranks might be random. When the $H_0$ is rejected, it simply means that; there is a statistically significant difference in performance between the pairs of algorithms being compared across all the 50 datasets.

As an outcome of the statistical analysis of the performance of the proposed method (sBal) against the studied methods when using Friedman/Bergman methods with a significance level of $\alpha = 0.05$, we observe a significant difference in performance between sBal and all the other methods where the $H_0$ is rejected with an exception of Random Forest. This means that our proposed methods achieved better performance as compared to the other method except for RAMOBoost where performance was comparable.

**Table 3** Statistical results for data from Table 2—sBal versus other ensemble methods on 50 imbalanced datasets

| Comparison | | | $z = (R_0 - R_1)/SE$ | $p$ | Result |
|---|---|---|---|---|---|
| SMOTEBoost | Versus | sBal | 6.89736 | 0 | H0 Rejected |
| RF | Versus | SMOTEBost | 4.3745 | 0.000012 | H0 Rejected |
| **RUBoost** | **Versus** | **sBal** | **4.351354** | **0.000014** | **H0 Rejected** |
| **Bagging** | **Versus** | **sBal** | **4.096754** | **0.000042** | **H0 Rejected** |
| SMOTEBag | Versus | SMOTEBoost | 3.726426 | 0.000194 | H0 Rejected |
| SMOTEBoost | Versus | RAMOBoost | 3.471825 | 0.000517 | H0 Rejected |
| **RAMOBoost** | **Versus** | **sBal** | **3.425534** | **0.000614** | **H0 Rejected** |
| **SMOTEBag** | **Versus** | **sBal** | **3.170934** | **0.001519** | **H0 Rejected** |
| Bagging | Versus | SMOTEBoost | 2.800606 | 0.005101 | H0 Rejected |
| RUBoost | Versus | SMOTEBoost | 2.546005 | 0.010896 | H0 Accepted |
| **RF** | **Versus** | **sBal** | **2.52286** | **0.01164** | **H0 Accepted** |
| RF | Versus | RUBoost | 1.828495 | 0.067475 | H0 Accepted |
| Bagging | Versus | RF | 1.573894 | 0.115512 | H0 Accepted |
| RUBoost | Versus | SMOTEBag | 1.180421 | 0.237833 | H0 Accepted |
| Bagging | Versus | SMOTEBag | 0.92582 | 0.354539 | H0 Accepted |
| RUBoost | Versus | RAMOBoost | 0.92582 | 0.354539 | H0 Accepted |
| RF | Versus | RAMOBoost | 0.902675 | 0.366699 | H0 Accepted |
| Bagging | Versus | RAMOBoost | 0.67122 | 0.502081 | H0 Accepted |
| RF | Versus | SMOTEBag | 0.648074 | 0.516937 | H0 Accepted |
| SMOTEBag | Versus | RAMOBoost | 0.254601 | 0.799032 | H0 Accepted |
| Bagging | Versus | RUBoost | 0.254601 | 0.799032 | H0 Accepted |

## 6 Conclusion and Future Works

In this paper, we have proposed a data level preprocessing sampling technique for ensemble methods for dealing with the classification of binary class imbalance problems. Different from the existing traditional Boosting, Bagging, and sampling-based ensemble methods, our proposed method does not alter the original class distribution of the dataset and does not discard any information that may be very important during the training process or introduce any new data that might cause overfitting.

The proposed method attempted to address the class imbalance problem from the perspective of improving the performance of ensemble-based algorithms by introducing a new data sampling technique that first converts the imbalanced binary dataset into multiple balanced binary datasets without introducing new instances or discarding any original instances. This is achieved by randomly splitting the instances from the majority class into multiple bags, with each bag having a copy of all minority instances and preserving the balance within each bag. After that, a given classification base algorithm is induced on the multiple balanced bags. Finally, the classification results resulting from the different base algorithms are combined into an ensemble using the majority voting ensemble rule.

An empirical study has been carried out on 50 binary imbalanced real-world datasets. Furthermore, a non-parametric Friedman and Bergmann's post-hoc statistical test was conducted, at a significant level of $\alpha = 0.05$, to validate the findings in the study. The experimental results demonstrate that the proposed method performs significantly better than all the other methods studied, except for RAMOBoost where performance was comparable. The statistical validation shows a statistically significant difference in performance between sBal and SMOTEBoost, RUBoost, Bagging, RAMOBoost and SMOTEBaging except for Random Forest hence suggesting that sBal is a good method and performs more effectively than the existing traditional methods.

Future works can leverage this balancing scheme and aim at finding a way to empower it with the ability to specify/configure the number of bags rather than the current situation where the number is dictated by the imbalance ratio of the dataset. We plan to influence the process of picking majority instances into the bags using pre-calculated data complexity measures, such as Instance Hardness [29], and ensure that each bag contains a mixture of instances with varying degrees of hardness. Unlike most bagging methods that use a uniform probability to select instances, this will ensure that base algorithms are trained on balanced bags, and at the same time containing diverse levels of hardness to learn various patterns in the datasets. This will further allow us to increase the number of base estimators similarly to other ensemble methods.

## References


1. Kaur H, Pannu HS, Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv 52(4)
2. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N (2018) A survey on addressing high-class imbalance in big data. J Big Data 5(1)
3. Fujiwara K et al (2020) Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis 8:1–15
4. Kasem A, Ghaibeh AA, Moriguchi H (2017) Empirical study of sampling methods for classification in imbalanced clinical datasets, vol 532
5. Tapkan P, Özbakir L, Kulluk S, Baykaso̧lu A (2016) A cost-sensitive classification algorithm: BEE-Miner. Knowl Based Syst 95:99–113
6. Jurek A, Bi Y, Wu S, Nugent C (2013) A survey of commonly used ensemble-based classification techniques. Knowl Eng Rev 29(5):551–581
7. Khoshgoftaar TM, Fazelpour A, Dittman DJ, Napolitano A (2016) Ensemble vs. data sampling: which option is best suited to improve classification performance of imbalanced bioinformatics data? In: Proceedings of the international conference on tools with artificial intelligence, ICTAI, vol 2016-January, pp 705–712
8. Alcalá-Fdez J et al (2011) KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Multi-Valued Logic Soft Comput 17(2–3):255–287
9. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C Appl Rev 42(4):463–484
10. Ye X, Li H, Imakura A, Sakurai T (2020) An oversampling framework for imbalanced classification based on Laplacian Eigenmaps. Neurocomputing 399:107–116
11. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2010) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern Part A Syst Hum 40(1):185–197
12. Zheng Z, Cai Y, Li Y (2015) Oversampling method for imbalanced classification. Comput Inf 34:1017–1037
13. Liu XY, Wu J, Zhou ZH (2006) Exploratory under-sampling for class-imbalance learning. In: Proceedings of the IEEE international conference on data mining, ICDM, pp 965–969
14. Hoens TR, Chawla NV (2013) Imbalanced datasets: from sampling to classifiers. In: Imbalanced learning: algorithms and applications, pp 43–59
15. Japkowicz N (2000) The class imbalance problem: significance and strategies. In: Proceedings. 2000 International conference on artificial intelligence, pp 111–117
16. Dittman DJ, Khoshgoftaar TM, Wald R, Napolitano A (2014) Comparison of data sampling approaches for imbalanced bioinformatics data. In: Proceedings of the 27th international Florida artificial intelligence research society conference, FLAIRS 2014, pp 268–271
17. Seiffert C, Khoshgoftaar TM, Van Hulse J (2009) Improving software-quality predictions with data sampling and boosting. IEEE Trans Syst Man Cybern Part A Syst Hum 39(6):1283–1294
18. Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) SMOTEBoost: improving prediction of the minority class in boosting, pp 107–119
19. Wei Q, Dunbrack RL (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 8(7)
20. Bi J, Zhang C (2018) An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. Knowl Based Syst 158(May):81–93
21. Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17(3):299–310
22. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One 12(6):1–17

23. Halimu C, Kasem A, Shah N (2019) Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In: International conference on machine learning soft computing, no. Mcc, pp 10–15
24. Oshiro TM, Perez PS, Baranauskas JA (2012) How many trees in a random forest? In: Lecture notes in computer science (including subseries, Lecture notes in artificial intelligence and Lecture notes in bioinformatics) LNAI, vol 7376, pp 154–168
25. Chen S, He H, Garcia EA (2010) RAMOBoost: Ranked minority oversampling in boosting. IEEE Trans Neural Networks 21(10):1624–1642
26. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
27. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701
28. Bergmann B, Hommel G (1988) Improvements of general multiple test procedures for redundant systems of hypotheses, no 1987, pp 100–115
29. Smith MR, Martinez T, Giraud-Carrier C (2014) An instance level analysis of data complexity. Mach Learn 95(2):225–256

# DyslexiAR: Augmented Reality Game Based Learning on Reading, Spelling and Numbers for Dyslexia User's


**Ibrahim Ahmad, Aza Jaiza Mohamad, Farah Farhana Roszali, and Norziah Sarudin**

**Abstract** Augmented Reality Game for based learning has been enhanced the learning experienced and developed the knowledge and skills of the user. The project methodology used in this study is the game development life cycle (GDLC). It includes initiation, pre-production, production, testing, beta testing and release. The purpose of this project is to produce video games for children with dyslexia who have visual and auditory learning difficulties related to memory, time management, speed processing, organization, organization and planning. The objective of this product was to develop Augmented Reality games for dyslexia students, second was to develop Reality-based games on reading, spelling and numbers for dyslexia students using the Unity Game Engine, the third was to test the appropriateness of learning based on reading, spelling games and numbers for dyslexic students. The number of user targets used to test this product are dyslexia students, teachers who teach dyslexia students, expert programmers and designer games and evaluators from the eLearning Carnival & Conference (eLCC 2019). The result of this DyslexiAR learning game is that dyslexic students have a better understanding of learning to read, spell and learn numbers.

**Keywords** Augmented reality · Game based learning · Dyslexia

I. Ahmad (✉) · A. J. Mohamad · F. F. Roszali · N. Sarudin
Faculty of Information and Communication Technology, Center of Advanced Computing Technology (C-ACT), Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Malacca, Malaysia
e-mail: ibrahim@utem.edu.my

A. J. Mohamad
e-mail: aza9590@gmail.com

F. F. Roszali
e-mail: farhanaroszali@gmail.com

N. Sarudin
e-mail: hajjahnorzie@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_22

## 1 Introduction

Dyslexia is defined as a difficulty with word level reading and spelling skills, which are in turn caused by phonological deficits [1]. In general, people with dyslexia appears very intelligent, but they are unable to read, write and spell in a correct way. Children with dyslexia have difficulties in learning how to read and write [2]. Rauschenberger et al. Has add that they are often diagnosed after they fail in school, even though dyslexia is not related to general intelligence. Students suffering from Dyslexia took more time than their peers to understand as well as to complete the task [3]. Some dyslexic children have difficulty in learning to read because they could not acquire the auditory equivalents of the appearances of the letters [4, 5]. Other than that, children with dyslexia are commonly associated with gross motor difficulties [6]. They are also have a problem with their motor skills that make them experience an issue of extraordinary pencil grasp and composing or replicate words [7]. Dyslexic person also has a poor memory, difficulty coloring and difficulty forming letters [8, 9].

Difficulties in spelling inconsistent conventions of the writing systems are common in children with dyslexia [10]. Individuals with dyslexia often perform even worse in spelling than in reading [11]. There might have spelled disability when dyslexic person is confusing with the sequence of letters in words. They also have a big chance of making mistakes when they put letter the wrong way around. For example, dyslexic person will write "b" rather than "d". They are confused between two letters that have the same writing technique.

Dyslexia for numbers or known as dyscalculia. Dyscalculia is a specific learning disability that related to math and the term Dyscalculia is created for the disability of performing mathematic operations [12]. However, dyscalculia can be educated and reduce the severity of the learning disability, however, never totally recovered [13]. Web technologies can help people with dyslexia to improve their reading and writing experience on the web [14]. Therefore, in this study, we propose an Augmented Reality Game technology based learning in reading, spelling and numbers for dyslexia. Augmented Reality Gamed offers a number of advantages of the present generation of learners [15]. This project improves their skills in reading, spelling and decoding numbers.

## 2 Methodology

The methodology use for this project is Game Development life Cycle (GDLC). It is consists of six phases as shown in Fig. 1 [16].

**Phase 1: Initiation**
The initial step to do in making a game is to draft a rough idea what sort of game that will be made. The output of initiation is the game idea and a basic game description.


**Fig. 1** Game development life cycle (GDLC)

Decide type of game to be developed, the target player and the objective of developing the game.

**Phase 2: Pre-production**
Pre-production includes the creation and the correction of game plan and the production of game model. Decide game engine to be use and develop game prototype outline and the concept.

**Phase 3: Production**
Production phase is the most important part which involved the creation of model and coding implementation. The prototype of each element needs to be creates in detail with more complete game mechanic and game element. Use prototype made in pre-production stage to develop the game.

**Phase 4: Testing**
Conduct an internal testing to detect bugs, error and glitches. Test the game playability and usability. The result of testing will decide whether it is time to continue the phase to the next one or need to go back to production cycle.

**Phase 5: Beta**
Conducted by the third party or game tester to test the game bug, error and glitches. The game is completed and finalized, need balancing and fixing if any problems occur. The output of beta testing is game tester will make a report of bugs and user feedback.

**Phase 6: Release**
Release is the final stage of game development life cycle. This stage involves the project documentation, product launching, planning for improvement and maintenance. After development completed, the game is ready to be publish to public.

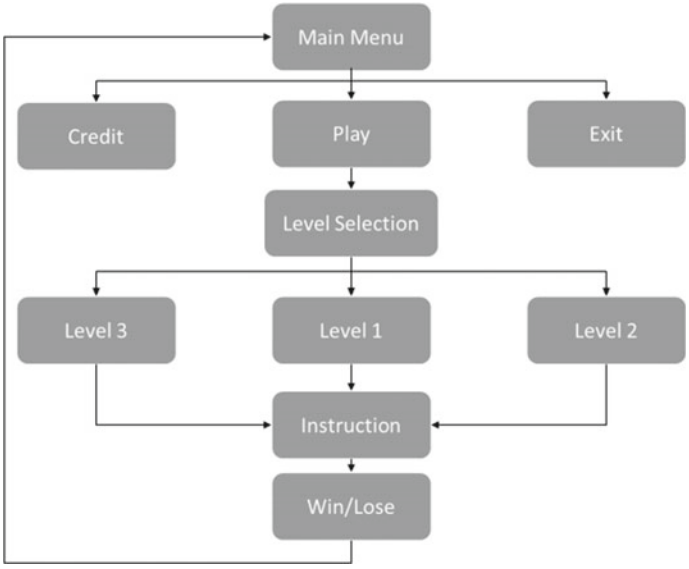



**Fig. 2** Game structure

## 3 System Overview

### *3.1 Game Structure*

In Fig. 2, the game structure shows that DyslexiAR will start with Main Menu. There are three options available in the Main Menu which are Credit to view the game credit, Play to start the game and Exit to quit the game. Level Selection will be available when the Play option is selected. There will be three options in the Level Selection which are Level 1, Level 2 and Level 3 to show there are three game levels available. Every single level selection, there will be an instruction on how to complete the level respectively. The game will then determine whether the level is completed with achieving objectives or vice versa to ascertain it is a win or lose condition. The game will then be directed back to The Main Menu.

### *3.2 Gameplay*

At the beginning of the game, the player will see the main menu. The main menu consists of three buttons which are Play, Credit and Exit. When player, click on Play button, the game will be directed to Level Selection as shown in Fig. 3.


**Fig. 3** Game level selection

Each level has different objectives for player to achieve. In Level 1, the player need to collect three cat and avoid dog within 60 s in order to win this level and proceed to next Level, which is Level 2 (Figs. 4 and 5).

To indicate the player has complete all of the game level, the All Level Completed Menu will be displayed as shown in Fig. 6.

If the player fail to complete a game level, the Game Over Menu will be displayed as shown in Fig. 7.


**Fig. 4** Level 1 instruction


**Fig. 5** Level 1 gameplay

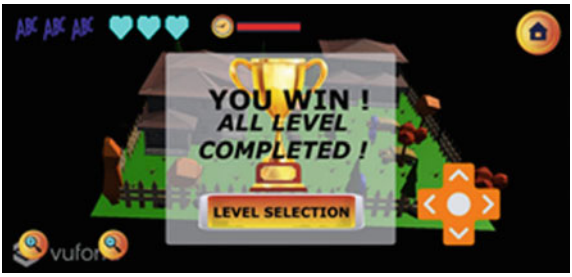



**Fig. 6** All level completed menu

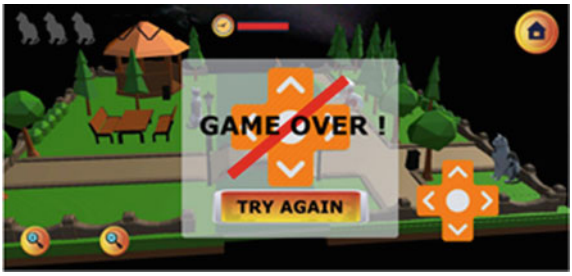



**Fig. 7** Game over menu

## 4 Result and Discussion

A playtest has been conducted on three different environment which are the company of Game Taiko Sdn Bhd, Fakulti Teknologi Kejuruteraan (FTK) and Sekolah Kebangsaan Padang Temu (SKPT). The three environment has three session which are Introduction session, Testing Game Session and Feedback session, except in the environment of Sekolah Kebangsaan Padang Temu which has four session that end with Questionnaire session.

The feedback collected from Game Taiko Sdn Bhd and FTK will be included in Conclusion and Future Scope since it is reviewed by Game Programmer, Game Designer and Evaluator respectively. The result of the playtest from SKPT are recorded.

### *4.1 Playtesting*

There is one set of questionnaire and simple quiz provided to the tester. Firstly, to identify the problem of the dyslexia student, a set of questionnaires is given to them. Secondly, one set of simple quiz is issued to the tester before and after they have

played the game. The same questionnaire as in before the tester playtest is given to the tester as to get the aftermath of playtesting the game.

**Before Playtest**

The first task to be given to the tester in a questionnaire is for them to write their own name as shown in Fig. 8. This test is to identify whether they can write their own name or not.

The first question in the quiz is provided to tester as in Fig. 9. This question is based on the picture of a cat. It shows that 1 out of 4 testers chose the wrong answer which is CAP. This is because student might confuse with letter P and T. Follow by 3 out of 4 testers chose the right answer which is CAT. This is because can spell simple words rather than longer word.

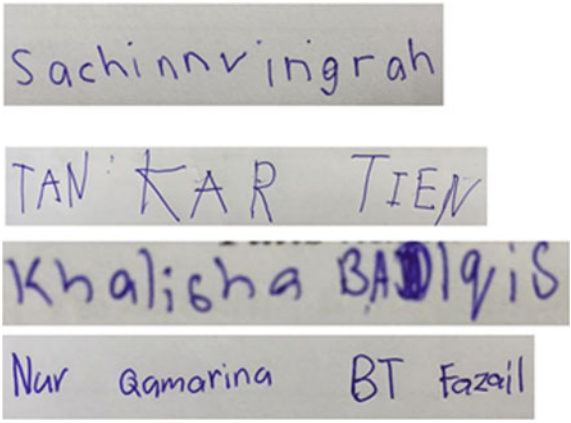



**Fig. 8** Name of Dyslexia tester

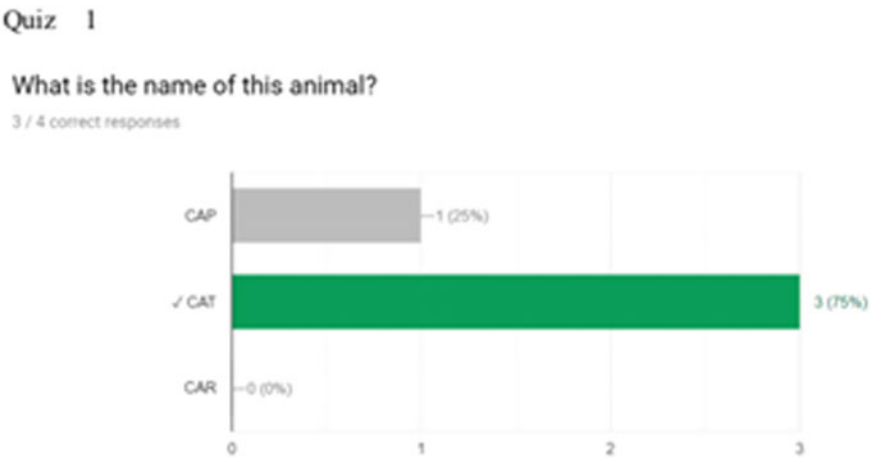



**Fig. 9** Quiz 1

The second question in the quiz is provided to tester as in Fig. 10. This question is based on the picture of a hat. It shows that 1 out of 4 testers chose the wrong answer which is black. This is due to the colour that might confuse them. Followed by 3 out of 4 testers chose a correct answer which is blue. This student can recognize the colour even though they did not play the game yet.

The third question in the quiz is provided to tester as in Fig. 11. This question is based on the picture of a number. It shows that 2 out of 2 testers chose the wrong answer which is option 2 (alphabet). This is because they have problem in replicate words or letter. They might be confused with the numbers and letter that have the same colour. Then, 2 out of 2 testers chose a correct answer as they know the differences of number and letters.

**After Playtest**

The same questionnaire as in before the tester playtest is given to the tester as to get the aftermath of playtesting the game. Figure 12 shows that 1 out of 4 testers chose a wrong answer which is CAP. This is because student might confuse with letter P and T. Follow by 3 out of 4 testers chose the right answer which is CAT. There are no changes of chart because the tester that still cannot remember the same things even though they have seen it.

The second question in the quiz is provided to tester as in Fig. 13. This question is based on the picture of a hat. It shows 4 out of 4 testers chose the correct answer which is blue. This game application helps them to recognize color correctly and make them focus and think before deciding.

The third question in the quiz is provided to tester as in Fig. 14. This question is based on the picture of a number. It shows 4 out of 4 testers chose the correct answer. It shows that this game application helps them to memorize things.

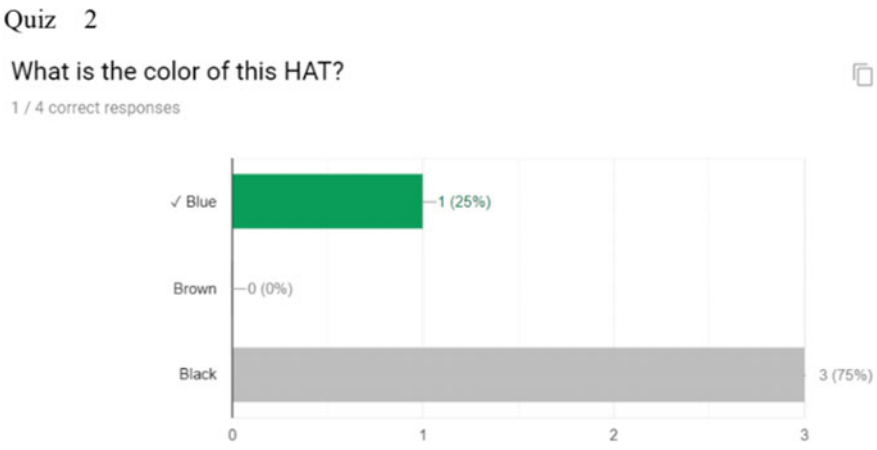



**Fig. 10** Quiz 2

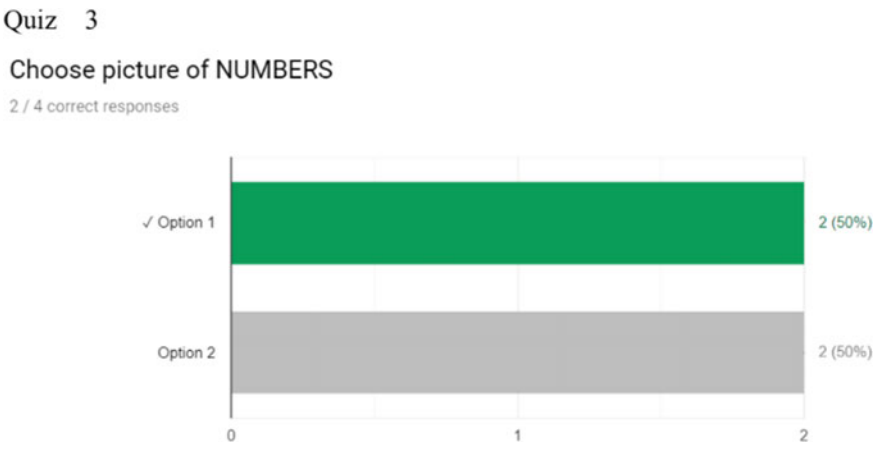



**Fig. 11** Quiz 3

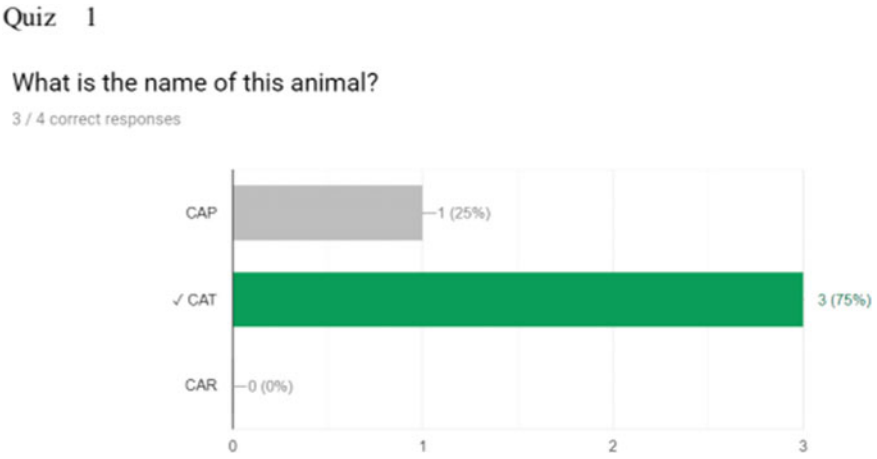



**Fig. 12** Quiz 1

## 5 Conclusion and Future Scope

In conclusion, DyslexiAR is suitable for students with dyslexia issue to enhance their learning skill. The Game Programmer and Game Designer from the company of Game Taiko Sdn. Bhd also has approved of this statement. To keep the player engaged in the game, the Evaluator from FTK has recommended to increase the in game rewards. Although the game is lacking in rewards, the tester keeps on playing the game more than one time as they feel amazed when the object appeared on screen, but there is nothing come out in real life. This result is a proof that the game currently

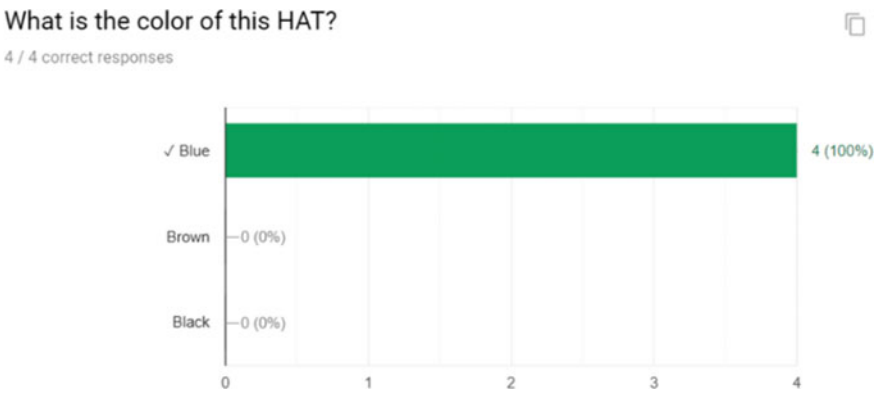



**Fig. 13** Quiz 2


**Fig. 14** Quiz 3

is fun as it is. In future, the game will be implemented with various challenges to keep the game immersive for players to enhance more of their learning skills.

**Acknowledgements** We would like to express our gratitude to Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Higher Education (MOHE), for the support of research grant FRGS/2018/FTMK-CACT/F00389 in this research.

## References

1. Adlof SM, Hogan TP (2018) Understanding dyslexia in the context of developmental language disorders. Lang Speech Hear Serv Sch 49(4):762–773
2. Rauschenberger M (2019) Early screening of dyslexia using a language-independent content game and machine learning

3. Kalsoom T et al (2020) Dyslexia as a learning disability: teachers' perceptions and practices at school level 42(1)155–166
4. Friedmann N, Coltheart M (2016) Types of developmental dyslexia, pp 1–37
5. Perry C, Zorzi M, Ziegler JC (2019) Understanding dyslexia through personalized large-scale computational models 30(3):386–395
6. Baharudin NS et al (2019) Gross motor skills performance in children with dyslexia: a comparison between younger and older children. J Sains Kesihatan Malaysia 17(2)
7. Hulme C, Snowling MJ (2016) Reading disorders and dyslexia 28(6):731
8. Reid G, Dyslexia: a practitioner's handbook. Wiley, Hoboken
9. Roitsch J, Watson S (2019) An overview of dyslexia: definition, characteristics, assessment, identification, and intervention. Sci J Educ 7(4)
10. Casalis S (2014) Written spelling in French children with dyslexia
11. Quemart P, Casalis S (2017) Morphology and spelling in French students with dyslexia: the case of silent final letters. Ann Dyslexia 67(1):85–98
12. Ariffin MM, Halim F, Abd N (2017) Mobile application for dyscalculia children in Malaysia. In: Proceedings of the 6th international conference on computing & informatics
13. Ferraz F, Neves J (2015) A brief look into dyscalculia and supportive tools. In: 2015 E-health and bioengineering conference (EHB). IEEE
14. Rauschenberger M, Baeza-Yates R, Rello L (2019) Technologies for dyslexia. In: Web accessibility. Springer, Berlin, pp 603–627
15. Das P, Zhu MO, McLaughlin L, Bilgrami Z, Milanaik RL (2017) Augmented reality video games: new possibilities and implications for children and adolescents. Multimodal Technol Interact 1(2)
16. Ramadan R, Widyani Y (2013) Game development life cycle guidelines. In: 2013 International conference on advanced computer science and information systems (ICACSIS). IEEE

# Applying Transfer Learning in Stock Prediction Based on Financial News


**Hai V. Che, Trung Q. D. Tran, and Duc M. Duong**

**Abstract** The most derived method and realistic way to predict the current stock price is via media resource and trusted new. In this paper, we will apply the current classifier text technique (Based LSTM) and pre-trained model from transfer learning to gain more intuition in financial news and precisely predict stock price. Finally, after using the latest pre-trained word embedding and a classification layer. We have achieved the robust success, and the experiment result shows that our method is able to outperform in accuracy than the previous one and have some advantage in the adaptive dataset.

**Keywords** Stock prediction · LSTM · ELMo · Transfer Learning

## 1 Introduction

The ubiquity of data today enables investors at any scale to make better investment decisions. The challenge is ingesting and interpreting the data to determine which data is useful, finding the signal in this sea of information. For this such classification task, machine learning and deep learning were used widely and gain more significant result lately. However, both these methods have their disadvantage, and most researchers in this field overlook this task and only focus on feature engineering and finance formula.

The recent progress of NLP and especially in deep learning field has shown an impressive result that we can achieve with pre-trained model and even possibly succeed the speed to maintain a real-time predicting system. Even for word embedding, the most successful work [1] has stated that the contextual embedding is more superior than the traditional one.

H. V. Che · D. M. Duong (✉)
University of Information Technology VNU-HCM, Ho Chi Minh, Vietnam
e-mail: ducdm@uit.edu.vn

T. Q. D. Tran
University of Technology VNU-HCM, Ho Chi Minh, Vietnam



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_23

In this paper, we only focus on daily one day ahead prediction and roughly split into three major parts, financial news pre-processing, feature engineering, and final prediction model. We propose to use the ELMo contextual language embedding [1] and [2] to automatically focus on the words that have the most impact on classification. The experiment result has shown the significant effect of our model, real-time prediction and can automatically enhance without being outdated.

## 2 Our Approach

For the recent year, the NLP community have been an inflection point for machine learning models handling text. Especially in the transfer learning scenario, we have had ELMO [7]. A new technique for embedding word into real vector space based on the context rather than a fix retrained weight for each token that was proposed in paper [6]. Besides the model surpasses the previous benchmark, using ELMO [7] as a pre-trained embedding for sentiment classification allow for a potential boost in performance.

### *2.1 System Design*

See Fig. 1.
Included:

- Using nlp to extract relevant stock within financial news
- Labeling data with specific strategies,
- Pre-process content and normalize,
- Word-embedding with fine-tune transfer learning,
- Training and evaluation state.

## 3 Processing Data

### *3.1 Crawling Data*

We've used MongoDB and Python script to automate crawling data from the multiple trusted source (Reuters and Bloomberg) that was in the required date (from October 2006 to December 2013) to match with the dataset from the confront paper [5]. And for stock historical data, we take the S&P 500 raw stock price data from the stock exchange and the stock prices are downloaded as a ".csv" file. Each line represents the

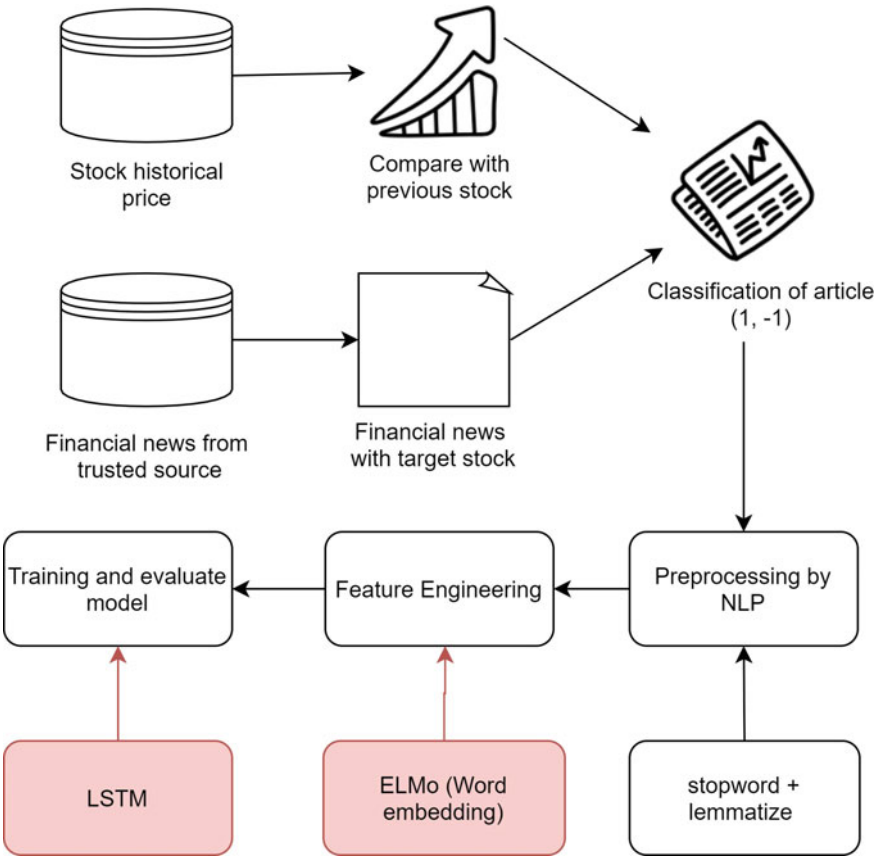



**Fig. 1** Pipeline system design

price of stocks in the S&P 500 by day including "$< id >$, $< symbol >$, $< date >$, $< open >$, $< high >$, $< low >$, $< close >$, $< volume >$ …". In it, each item has the following meanings: trading code, stock code, trading date, opening price, highest price, lowest price, closing prices, trading volume, …etc

## *3.2 Preprocess and Labeling Articles*

For the intuitive comparison with the paper [5], we will use the same labeling technique and grouping method with some minor change.

All information collected is tokenize and normalize by the Spacy library [3], with high precision in English word. For the financial news that we have collected, all heading and body of story were used and split them into sentences. We only keep

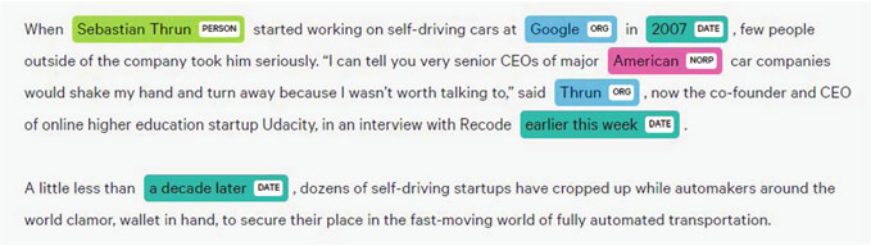



**Fig. 2** Entity extraction

those sentence that mentions at least a stock names or a public company. Not only the simple search method was used to find the stock name, but the **entity extraction** method was also used to finding the most accuracy sentence that mentions the stock (Fig. 2).

For the sentence that has multiple stocks, each mention stock will be labeled by the same sentence. Then, we group all these sentences by the publication dates and the names of the underlying stock to form the samples. Each sample contains the sentences that were published on the same date and mentions the same stock and company. Moreover, the sample was label as positive and negative (price up or down) based on the next day's closing price consulted from the dataset. For example: +1 corresponds to the trend increase of the stock price, −1 corresponds to the downtrend of the stock price.

## *3.3 Dataset Splitting*

The data set will be randomly segment into three sets: train, dev and test; at the rate of **0.75**| **0.125**||**0.125** as in the article [5]. Based on the dataset describe in the above paper. We split the dataset into three sets as follows (Table 1).

The set of train and valid will be put into training and embedding word, the test set will be evaluated at the experimental step. The data is divided as above to get trusted results, in order to help the model be more accurately evaluated with the experimental data set that has never been seen before. Total data negatives and positives are also collected in a balanced manner (**42674 positive** and **41531 negative**).

**Table 1** Dataset details

| Dataset | Date publish | Data samples |
|---|---|---|
| Train | 2006-10-01 to 2012-12-31 | 63,153 |
| Valid | 2013-01-01 to 2013-06-15 | 10,525 |
| Test | 2013-06-16 to 2013-12-31 | 10,527 |

## 4 Word Embedding

The feature engineering task here will only be using ELMo embedding [7]. This simplicity will enhance the performance in extracting feature and will not be outdated if the domain word is changing.

The task-specific part is the combination of the task-agnostic representations. Concretely, in ELMo, the word representation is computed with the following equation:

$$ELMo_k^{task} = \epsilon(\mathbb{R}_k; \theta^{task}) = \gamma \sum_j^L s_j^{task} h_{kj}^{lm} \quad (1)$$

The indices $k$ and $j$ each called to the index of the word and the index of the layer that the feature is being extracted from language model. More specifically of $h_{kj}$ is the output of the $j_{t,h}$ LSTM layer for the word $k$. $j$ is the weight of $h_{kj}$ in computing the representation for $k$ that are softmax-normalized. The parameter $\gamma$ is a dependent value that allows for scaling the entire vector, which is important during optimization.

The reason that ELMo is transfer learning is this simple, as the task is taken input as a discrete token, meaning ELMo can be used with preexisting embeddings or can replace them and be trained end-to-end. Despite the fixed hidden states, the model still has the flexibility to feature various layers of abstraction depending on the context of the input. The ELMo pretrain we used in this paper is taken from https://tfhub.dev/google/elmo/2Tfhub.

## 5 Deep Neural Network

### *5.1 Model*

The structure of the model is the basic LSTM [2] for the classified text task. As the Lstm layer takes a lot of time to converge, the model consisted of only one Lstm layer and multiple fully- connected layer. Cause the labeled data was an only negative or positive value; the output layer is a sigmoid function to compute posterior probabilities of the last hidden layer. The structure of the model will be expressed as follows (Table 2):

With the following configuration:

- batch_size: 32
- epoch: 50
- learning_rate (Adam optimizer): 0.001.

**Table 2** Model layer params

| Layer (type) | Description | Param (units) |
|---|---|---|
| Word_Embedding | Size of word embedding | 128 |
| Elmo_Embedding | Size of word Elmo embedding | 1024 |
| Concatenate_layer | Concentrate two embedding layer | 1152 |
| Batch_norm | Normalize layer | 0 |
| Bi_lstm | Bidirectional LSTM layer | 100 |
| Dense_layer | Hidden layer | 1 |
| Binary_crossentropy | Evaluate layer | 1 |

### *5.2 Optimization*

For the optimization task, we simply use Batch normalization [4] and Lasso regularization [8] for avoiding overfitting and make the training more faster. The Batch normalization can help us stabilizing the distribution of layer before activations throughout training, reducing the instability of deeper neural nets to saturate or diverge. But the Batch normalization only have a little effect in avoiding overfitting, thus the Lasso technique was used to adding weight decay, that every time we update a weight $w$ with the gradient $\Delta j$ in respect to $w$, we also subtract from it $\gamma * w$. This method gives the weights a tendency to decay towards zero, increasing error in the training process.

In order to avoid overfitting in deep learning model, the EarlyStopping technique is used with the parameter used to stop training is accuracy of the valid set (valid accuracy). Using Early Stopping to prevent too long a model training leads to overfitting. So each epoch if the valid accuracy does not increase it will stop training and save the train weight.

## 6 Experiment Result

The purpose is to find the most optimal model for the stock market trend prediction system. We gave the data set to be trained in many different models to show the visualization of the training results and to compare the model in paper [5]. All models are tested on test set sample (Table 3)

The training process was built on a combination of algorithms, because of the large data source, the average epoch consumed 10 min. Here is the training process of the models that have been saved and represented in graph form (Fig. 3).

As the graph shows, other embedding techniques such as Glove [6] did not perform as well as ELMo transfer embedded techniques. All models are applied with

**Table 3** Result compare

| Feature combination | Accuracy rate |
|---|---|
| Glove + Dense | 0.51 |
| ELMo + Dense | 0.50650 |
| Plain lstm | 0.55 |
| Glove + Lstm | 0.52821 |
| Leverage financial news | 0.5435 |
| ELMo + Lstm | 0.56542 |

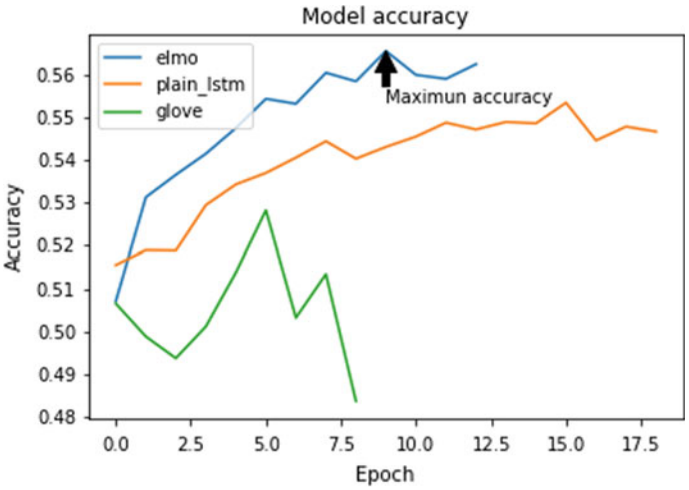



**Fig. 3** Learning curve comparison

EarlyStopping technique to limit the calculation of redundancy. By checking with the test data set that has not been seen, we can see positive results in accurately predicting the standard trend of stock prices.

## *6.1 Practical Experiment*

In addition, we also checked the prediction of a number of individual stocks out of 500 stocks appearing in the S&P 500. We have selected five stocks with high trading volume: SHOP, MELI. , GDOT, BBY, C. In the period from May 15, 2016 to May 29, 2016 not mentioned in the data (14 days) (Table 4).

We have collected over 20 news related to target stock and dates. The trend will be modeled according to each news item per day (1 is equal to **Uptrend** and −1 is **Downtrend**) and summed up as the predicted indicators. If the value is greater than

**Table 4** Liquidity of compare stocks

| TT | Stock Code | Maximum trading volume | Minimum trading volume | Average trading volume |
|---|---|---|---|---|
| 1 | SHOP | 10,897,100 | 744,200 | 636,796 |
| 1 | SHOP | 10,897,100 | 744,200 | 636,796 |
| 2 | MELI | 14,763,506 | 2,575,451 | 8,669,478 |
| 3 | GDOT | 81,370,241 | 17,979,893 | 49,675,067 |
| 4 | BBY | 7,506,500 | 281,200 | 3,893,850 |
| 5 | C | 32,123,22 | 939,283 | 2,075,802 |

**Table 5** Predicting results of 5 stocks in 14 days

| | SHOP | | | MELI | | | GDOT | | | BBV | | | C | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Open | Close | Trend | Open | Close | Trend | Open | Close | Trend | Open | Close | Trend | Open | Close | Trend |
| 15/5 | 45.5 | 44 | 0 | 33.6 | 33.6 | 1 | 7.0 | 7.1 | 0 | 12.5 | 12.4 | 0 | 45.3 | 45.3 | 1 |
| 16/5 | 42.9 | 40.3 | 0 | 33.5 | 34.3 | 1 | 7.3 | 7 | 1 | 14 | 13.8 | 0 | 45.7 | 45.8 | 1 |
| 17/5 | 41.3 | 44 | 0 | 33.8 | 32 | 1 | 6.9 | 7.4 | 1 | 14.1 | 13.2 | 1 | 45 | 44.6 | 1 |
| 18/5 | 42.5 | 40.2 | 1 | 32.1 | 30.9 | 1 | 7.1 | 6.8 | 1 | 13.2 | 13.9 | 1 | 44.5 | 44.7 | 1 |
| 19/5 | 40.1 | 41 | 1 | 31.1 | 31.3 | 0 | 6.6 | 7.2 | 1 | 14.4 | 14.4 | 1 | 44.9 | 45.2 | 1 |
| 20/5 | 40.8 | 41.1 | 1 | 30.7 | 31.7 | 1 | 7.1 | 7 | 0 | 14.4 | 14.2 | 0 | 45.2 | 46.1 | 0 |
| 21/5 | 41 | 42.1 | 1 | 32.1 | 31.3 | 0 | 7 | 7.8 | 1 | 14.3 | 14.5 | 1 | 45.6 | 45 | 0 |
| 22/5 | 42 | 42.19 | 0 | 31.8 | 32.4 | 1 | 7.9 | 8.45 | 0 | 14.5 | 15.02 | 0 | 44.7 | 44.8 | 1 |
| 23/5 | 42.1 | 42.9 | 1 | 32.2 | 31.7 | 1 | 8.7 | 8.7 | 1 | 15.2 | 15.1 | 0 | 44.7 | 44.0 | 1 |
| 24/5 | 43.7 | 44.1 | 1 | 32.2 | 33.3 | 0 | 8.7 | 8.83 | 1 | 15.1 | 15 | 0 | 43.9 | 44.1 | 0 |
| 25/5 | 44.9 | 44.5 | 0 | 34.2 | 36.4 | 1 | 8.6 | 8.6 | 0 | 14.9 | 15 | 0 | 44 | 43.5 | 0 |
| 26/5 | 45.2 | 45.7 | 1 | 36 | 36.4 | 1 | 8.8 | 9.4 | 0 | 15.0 | 14.8 | 0 | 43.5 | 42.7 | 0 |
| 27/5 | 45.7 | 45.2 | 0 | 35.7 | 34.5 | 0 | 9.5 | 9.2 | 0 | 14.7 | 14.5 | 0 | 42.8 | 42.4 | 0 |
| 28/5 | 45.5 | 44 | 0 | 34.8 | 33.7 | 1 | 8.8 | 8.5 | 0 | 14.5 | 13 | 0 | 42.3 | 42.5 | 0 |
| 29/5 | 42.9 | 40.3 | 0 | 33.6 | 33.6 | 0 | 8.8 | 8.3 | 0 | 12.5 | 12.4 | 0 | 45.3 | 45.3 | 1 |

or equal to 0 then it is 1(Uptrend), the smaller is 0(Downtrend) in that day. Opening and closing numbers in stock codes will be rounded to one unit. The predict trend will be evaluated with actual trend and the result have been described as Table 5.

Included:

- 1: predict Uptrend.
- 0: predict Downtrend.
- Blue : Means predicting right trending.
- Red : Means predicting wrong trending.

As we can see in Table 5, the stocks codes C, GDOT and SHOP have an accuracy of nearly 57–58% but the remaining stocks have an accuracy of less than 54% but still better than the forecast randomly guess the price. By checking the above, the team found quite satisfactory results in predicting the trend of the stock moves based on stock prices.

## 7 Conclusion

In this paper, we have introduced a transfer learning approach for better gaining intuition from financial news in the stock price movement problem. This approach does not rely on domain knowledge and will not be limited by an obsolete traditional analyst. Our experiment has shown that our proposed method can improve the prediction accuracy on a confront dataset and have practical usage in real world.

## References

1. Clark K, Luong M, Manning CD, Le QV (2018) Semi-supervised sequence modeling with cross-view training. CoRR http://arxiv.org/abs/1809.08370
2. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:8. https://doi.org/10.1162/neco.1997.9.8.1735
3. Honnibal M, Montani I (2017) spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
4. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR http://arxiv.org/abs/1502.03167
5. Peng Y, Jiang H (2016) Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 374–379. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-1041, https://www.aclweb.org/anthology/N16-1041
6. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. EMNLP. 14:1532–1543
7. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long Papers). pp 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202, https://www.aclweb.org/anthology/N18-1202
8. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Statist Soc (Ser B) 58:267–288

# Solving Time-Fractional Parabolic Equations with the Four Point-HSEGKSOR Iteration


**Fatihah Anas Muhiddin, Jumat Sulaiman, and Andang Sunarto**

**Abstract** The goal is to show the usefulness of the 4-point half-sweep EGKSOR (4HSEGKSOR) iterative scheme by implementing the half-sweep approximation equation based on the Grünwald-type fractional derivative and implicit finite difference (IFD) method to solve one-dimensional (1D) time-fractional parabolic equations compared to full-sweep Kaudd Successive over-relaxation (FSKSOR) and half-sweep Kaudd Successive over-relaxation (HSKSOR) methods. The formulation and implementation of the 4HSEGKSOR, HSKSOR and FSKSOR methods are also presented. Some numerical tests were carried out to illustrate that the 4HSEGKSOR method is superior to HSKSOR and FSKSOR methods.

**Keywords** Grünwald derivative · 4HSEGKSOR iteration · Finite difference method · Time fractional parabolic equation

## 1 Introduction

In many fields of engineering, mathematical models are usually developed in the form of partial differential equations (PDEs). Due to its capability in describing materials with memory and hereditary properties, fractional-order PDEs (FPDEs) have earned great interests among many distinguished researchers in the field of fractional calculus. Wide applications of fractional calculus can be seen such as in biological cells and tissues model [1, 2], ecological model [3] and rheological polymer model [4]. Over the last few decades, there has been increasing interest

F. A. Muhiddin (✉)
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: fatihah.anas@uitm.edu.my

J. Sulaiman
Mathematics with Economics Program, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

A. Sunarto
Faculty of Economic and Islamic Business, IAIN, Bengkulu, Indonesia
e-mail: andang99@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_24

in studies on the time-fractional diffusion equations with the existence of a source function $f(x, t)$, e.g. papers by [5–7]. While, majority of the earlier studies have considered the homogeneous problem instead, i.e. when $f(x, t) = 0$.

Several numerical techniques of FPDEs were studied in earlier works for instance, the finite difference discretization scheme, the finite element, the finite volume and collocation methods [8–14]. The most commonly used definitions of fractional derivatives are based on the Riemann-Liouville and Caputo fractional derivatives. In this paper, we employed the Grünwald-type fractional derivative and the implicit discretization scheme for solving time-fractional parabolic equations (TFPEs). The occurrence of large linear systems (SLEs) that arise from the discretization of problem (1) usually require some iterative treatments.

To reduce the computational complexity of the standard iterative method, block-iterative method and half-sweep iteration concept has been implemented extensively in previous studies. The applications including but not limited to, solving the poisson image blending problem [15], robot path planning problem via nine-point Laplacian method [16], two-point fuzzy boundary problem [17], two-dimensional Helmholtz equation [18] and boundary value problems related to the ordinary fractional differential equations [19, 20].

Additionally, since it was first proposed by Youssef [21], the applications of KSOR method have been widely demonstrated through the application of the full-and half-sweep concept in solving various types of research problem such as the two-point boundary problem [22] and the Fredholm integral equations system of second kind [23]. By extending the previous study, this paper intended to evaluate the performance of the half-sweep concept when combined with the four-point EGKSOR. In this work, we named the proposed iterative scheme as 4HSEGKSOR.

To examine the performance of this proposed scheme, let us consider 1D time-fractional parabolic equations (TFPE) defined as

$$\frac{\partial^\alpha U}{\partial t^\alpha} + C_1(x)\frac{\partial U}{\partial x} + C_2(x)\frac{\partial^2 U}{\partial x^2} = f(x,t), \quad 0 \le x \le l, \quad t > 0, \quad \alpha \in (0,1), \quad (1)$$

with initial condition

$$U(x, 0) = g_1(x), \quad 0 \le x \le l,$$

and the boundary condition

$$U(0, t) = g_2(t); \quad U(l, t) = g_3(t), \quad 0 < t \le T$$

where $C_1(x)$ and $C_2(x)$ can be constants or function of $x$. Here, Eq. (1) is equivalent to the standard parabolic PDEs at $\alpha = 1$. To ensure simplicity, supposed that the solution domain (1) is in equal part subintervals ($N = 2^p$, $p \ge 2$), which noted as $\Delta x$ and $\Delta t$, corresponds to the directions of $x$ and $t$ respectively and where

$$\Delta x = h = \frac{l}{N}, \quad n = N - 1, \quad \Delta t = \frac{T}{M} \tag{2}$$

There are several numbers of fractional derivatives. One way to represent the discrete fractional derivatives is by Grünwald fractional operator, which the truncated Grünwald fractional derivative formula, $D_G^{\alpha}$ of order $\alpha$ for function $f(t)$ is defined as [24, 25]

$$D_G^{\alpha} f(t) = \frac{1}{(\Delta t)^{\alpha}} \lim_{N \to \infty} \sum_{\kappa=0}^{N} g_{\alpha,\kappa} f(t - \kappa \Delta t), \quad 0 < \alpha < 1 \tag{3}$$

and the normalized Grünwald-weights are given by

$$g_{\alpha,\kappa} = \frac{\Gamma(\kappa - \alpha)}{\Gamma(-\alpha)\Gamma(\kappa + 1)}, \tag{4}$$

where, by using the recurrence relationships

$$\begin{aligned} g_{\alpha,0} &= 1; \\ g_{\alpha,\kappa} &= \left(1 - \frac{\alpha + 1}{\kappa}\right) g_{\alpha,\kappa-1}, \quad \kappa = 1, 2, 3, \ldots \end{aligned} \tag{5}$$

Discussion on the next sections are summarized as follows. Section 2 discussed how the finite difference approximation equations of TFPEs were formulated based on the Grünwald-type derivative. After that, the derivation of 4HSEGKSOR iterative method is deliberated in Sect. 3. Next, the experimental results related to the tested numerical examples are analyzed and their efficiencies are discussed in Sect. 4. Meanwhile, the last Section concludes the study.

## 2 Half-Sweep Approximation Equation

Before problem (1) could be discretized, the solution domain is partitioned so that uniform finite grid-points are formed. This is done to facilitate the discretization process of the problem (1). To start, first the mesh point is stated as (Fig. 1) $x_i = \alpha + ih$, where $i \in [\phi, \eta]$ and

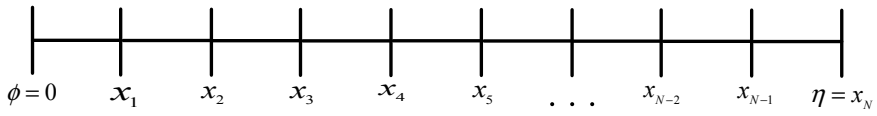



**Fig. 1** The equidistance interior node points

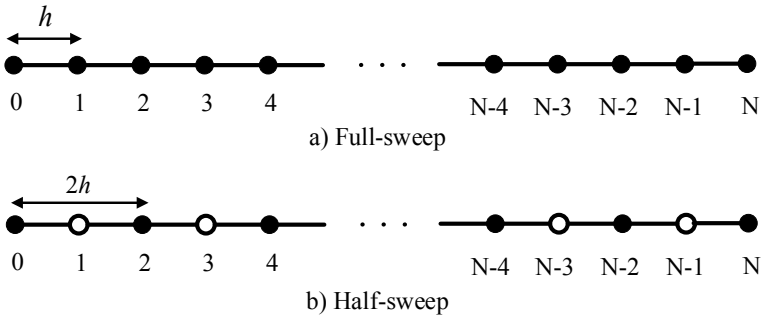



**Fig. 2** The uniform grid network distribution of **a** full-and **b** half-sweep cases

$$h = \Delta x = \frac{\eta - \phi}{N}, \quad n = N - 1 \tag{6}$$

As mention in Sect. 1, the purpose of imposing the half-sweep iterative scheme is to improve the convergence rate of an iterative method. Whereby, the length of each grid size is $2h$, compared to just single $h$ for the standard full-sweep concept. The grid point distributions of the full-and half-sweep cases are as explained in Fig. 2.

Next, to generate the finite difference approximation for problem (1), we substitute the time with Eq. (3), and central difference scheme to space. The derivation will end up with

$$\begin{aligned} &\frac{1}{(\Delta t)^{\alpha}} \sum_{k=0}^{j} g_{\alpha,\kappa} U_{i,j-k} + \frac{C_1(x)}{2p\Delta x}\left(U_{i+p,j} - U_{i-p,j}\right) \\ &\quad + \frac{C_2(x)}{(p\Delta x)^2}\left(U_{i+p,j} - 2U_{i,j} + U_{i-p,j}\right) = f_{i,j} \end{aligned} \tag{7}$$

where $p = 1$ for full-sweep iteration, while $p = 2$ for the half-sweep concept.

Hence, by letting

$$G_k = \frac{g_{\alpha,\kappa}}{(\Delta t)^{\alpha}}, \quad \rho_i = \frac{C_1(x)}{4\Delta x}, \quad \varphi_i = \frac{C_2(x)}{(2\Delta x)^2},$$

Equation (4) for the half-sweep cases can be simplified into

$$\alpha_i U_{i-2,j} + \beta_i U_{i,j} + \gamma_i U_{i+2,j} = F_{i,j} \tag{8}$$

where

$$F_{i,j} = \begin{cases} f_{i,1} - G_1 U_{i,0} & j = 1 \\ f_{i,j} - \sum_{\kappa=1}^{j} G_k U_{i,j-\kappa} & j = 2, 3, \ldots, M, \end{cases} \tag{9}$$

and

$$\alpha_i = \varphi_i - \rho_i, \quad \beta_i = G_0 - 2\varphi_i, \quad \gamma_i = \varphi_i + \rho_i. \tag{10}$$

Then, to facilitate us in solving the TFPE, Eqs. (5), (6) and (7) were used to reconstruct the SLEs in matrix form as

$$A\underline{U}_j = \underline{F}_j \tag{11}$$

where

$$A = \begin{bmatrix} \beta_2 & \gamma_2 & & & & & \\ \alpha_4 & \beta_4 & \gamma_4 & & & & \\ & \alpha_6 & \beta_6 & \gamma_6 & & & \\ & & \alpha_8 & \beta_8 & \gamma_8 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \alpha_{N-4} & \beta_{N-4} & \gamma_{N-4} \\ & & & & & \alpha_{N-2} & \beta_{N-2} \end{bmatrix}_{(\frac{N}{2}-1)\times(\frac{N}{2}-1)},$$

$$\underline{U}_j = \left[ U_{2,j} \; U_{4,j} \; U_{6,j} \cdots U_{N-4,j} \; U_{N-2,j} \right]^{\mathrm{T}},$$

$$\underline{F}_j = \left[ F_{2,j} - \alpha_2 U_{0,j} \; F_{4,j} \; F_{6,j} \cdots F_{N-4,j} \; F_{N-2,j} - \gamma_{N-2} U_{N,j} \right]^{\mathrm{T}}.$$

## 3 Derivation of the Iterative Method

Based on the half-sweep iterative scheme explained in the previous section and referring to the sparse SLE in Eq. (11), in this section, the formulation and implementation of the 4HSEGKSOR iterative method are discussed. To begin formulating the 4HSEGKSOR iterative method, we recall the formula of standard full-sweep SOR iterative method. For $A = D - L - V$ (D diagonal, L strictly lower, V strictly upper triangular), the SOR method in matrix form as defined by [26–28]

$$\underline{U}_{SOR}^{(\kappa+1)} = (D - \omega L)^{-1}[(1-\omega)D + \omega V]\underline{U}^{(\kappa)} + \omega(D - \omega L)^{-1}F, \tag{12}$$

and then refer to [21] for the full-sweep KSOR iterative methods written in matrix form as

$$\begin{aligned} \underline{U}_{KSOR}^{(\kappa+1)} = &\left[\left(1+\omega^*\right)D - \omega^* L\right]^{-1}\left(D + \omega^* V\right)\underline{U}^{(\kappa)} \\ &+ \left[\left(1+\omega^*\right)D - \omega^* L\right]^{-1}\left(\omega^* F\right) \end{aligned} \tag{13}$$

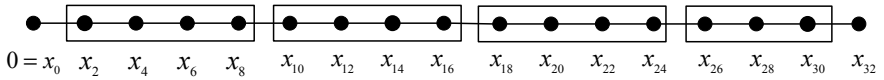



**Fig. 3** Implementation of the 4HSEGKSOR at the solution domain for m = 32

Here we wanted to highlight on the domain of relaxation factors $\omega^*$ of KSOR iterative method, which has been extended to $\omega^* \in R - [-2, 0]$, versus $\omega \in (0, 1]$ for the classical SOR iterative method.

Figure 3 shows how the block iterative methods were imposed on the half-sweep grid-point by grouping it into four points ($4 \times 4$) block each. Here, we can see that blocks of four-point and three-point are formed. The four-point HSEGKSOR iterative method is applied onto the four-point blocks, whereas the last three-point block is treated as an ungrouped case. This process is then iterated until it converges.

Hence, the formulation of 4HSEGKSOR method of the four-point blocks at $i = 2, 10, 18, \ldots, N - 14$, given in general form as

$$\begin{bmatrix} U_{i,j} \\ U_{i+2,j} \\ U_{i+4,j} \\ U_{i+6,j} \end{bmatrix}^{(\kappa+1)} = \frac{1}{(1+\omega^*)} \begin{bmatrix} U_{i,j} \\ U_{i+2,j} \\ U_{i+4,j} \\ U_{i+6,j} \end{bmatrix}^{(\kappa)} + \frac{\omega}{(1+\omega^*)}^* \begin{bmatrix} r_i & 1 & 0 & 0 \\ q_{i+2} & r_{i+2} & 1 & 0 \\ 0 & q_{i+4} & r_{i+4} & 1 \\ 0 & 0 & q_{i+6} & r_{i+6} \end{bmatrix}^{-1} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix} \tag{14}$$

where

$$S_1 = \mathrm{X}_{i,j} - q_i U_{i-2,j},\ S_2 = X_{i+2,j},$$
$$S_3 = X_{i+4,j},\qquad S_4 = X_{i+6,j} - U_{i+8,j}.$$

Meanwhile, for the ungrouped three-point block, the last block at $i = N - 6$ (3HSEGKSOR) method is given in general form as

$$\begin{bmatrix} U_{n-5,j} \\ U_{n-3,j} \\ U_{n-1,j} \end{bmatrix}^{(\kappa+1)} = \frac{1}{(1+\omega^*)} \begin{bmatrix} U_{n-5,j} \\ U_{n-3,j} \\ U_{n-1,j} \end{bmatrix}^{(\kappa)} + \frac{\omega^*}{(1+\omega^*)} \begin{bmatrix} r_{n-5} & 1 & 0 \\ q_{n-3} & r_{n-3} & 1 \\ 0 & q_{n-1} & r_{n-1} \end{bmatrix}^{-1} \begin{bmatrix} S_{n-5} \\ S_{n-3} \\ S_{n-1} \end{bmatrix} \tag{15}$$

where

$$S_{n-5} = X_{n-5,j} - q_{n-5}U_{n-7,j}, \quad S_{n-3} = X_{n-3,j}, \quad S_{n-1} = X_{n-1,j} - U_{n+1,j+1}.$$

while $q_i = \frac{\alpha_i}{\gamma_i}$, $r_i = \frac{\beta_i}{\gamma_i}$, and $\mathrm{X}_{i,j} = \frac{F_{i,j}}{\gamma_i}$.

Thus, Algorithm 1 gives the summary on the implementation of Eqs. (14) and (15) for the 4HSEGKSOR iterative method.

**Algorithm 1** 4HSEGKSOR scheme

1. Initialize $\underline{U}_j^{(0)} \leftarrow 0$ and $\varepsilon \leftarrow 10^{-10}$,
2. Identify the optimal weightage values of $\omega^*$,
3. For $i = 2, 10, 18, \ldots, N-14$, from Eq. (14) calculate the 4-point group, For $i = N-6$, from Eq. (15) calculate the ungroup case.
4. Perform the convergence test, $\left|\underline{U}_{i,j}^{(\kappa+1)} - \underline{U}_{i,j}^{(\kappa)}\right| \leq \varepsilon$. If yes, go proceed to step (5). Otherwise rerun step (3).
5. Display approximate solution.

## 4 Numerical Experiment

In this study, three numerical experiments of 1D time-fractional parabolic equations problem were considered for effectiveness verification of the proposed methods. As comparison purposes, three components have been considered, which are the number of iterations, time of execution and maximum absolute error to measure the performance of the 4HSEGKSOR over the other two KSOR point-iterative methods. Following are the examples, which the tolerance of error has been set to $\varepsilon = 10^{-10}$. The tested problems are considered in the following form:

$$\begin{cases} \frac{\partial^\alpha U(x,t)}{\partial t^\alpha} + C_1(x)\frac{\partial U(x,t)}{\partial x} + C_2(x)\frac{\partial^2 U(x,t)}{\partial x^2} = f(x,t), & t > 0, x \in R, \alpha \in (0,1), \\ u(x,0) = g_0(x), & 0 < x < 1 \\ u(0,t) = g_1(t), & t > 0, \\ u(1,t) = g_2(t), & t > 0, \end{cases} \tag{16}$$

**Example 1** Consider the TFPE problem (16) with $C_1(x) = 1$, $C_2(x) = -1$ and the following additional data [29]:

$$\begin{cases} f(x,t) = \frac{2t^{2-\alpha}}{\Gamma(3-\alpha)} + 2x - 2, \\ g_0(x) = x^2, \end{cases}$$

The exact solution is $u(x,t) = x^2 + t^2$.

**Example 2** Consider the TFPE problem (16) with $C_1(x) = 0$, $C_2(x) = -1$ and the following additional data [30]:

$$\begin{cases} f(x,t) = \frac{2t^2}{\Gamma(3-\alpha)t^\alpha} \sin(2\pi x) + \left(4\pi^2 t^2\right) \sin(2\pi x), \\ g_0(x) = 0, \end{cases}$$

The exact solution is $u(x,t) = t^2 \sin(2\pi x)$.

**Example 3** Consider the TFPE problem (16) with $C_1(x) = 0$, $C_2(x) = -1$ and the following additional data [31]:

$$\begin{cases} f(x,t) = \frac{2e^x t^2}{\Gamma(3-\alpha)t^\alpha} - e^x / t^{-2}, \\ g_0(x) = 0, \end{cases}$$

The exact solution is $u(x,t) = e^x / t^{-2}$.

Hence, based on five distinct mesh sizes (m = 128, 256, 512, 1024, 2048), Tables 1, 2 and 3 recorded the results of the 4HSEGKSOR method for the tested problems.

Based on the numerical outputs generated from the tested examples as depicted in Tables 1, 2 and 3, the improvement of the convergence rate after the application of 4HSEGKSOR method in contrast to the FSKSOR and HSKSOR point iterative methods are summarized in Tables 4 and 5.

## 5 Conclusion

In this work, we discussed the implementation of the 4HSEGKSOR method in solving the Grünwald implicit difference approximation equations. Observed from the experimental results, applying half-sweep iterative concept to the regular full-sweep KSOR iterative method has indeed improved the FSKSOR point iterative method. Meanwhile, the performance of HSKSOR iterative method has improved even further upon extended application of the block 4HSEGKSOR iterative method. This led to the conclusion, that the efficiency of 4HSEGKSOR iterative method is undoubtedly better, in terms of the number of iterations and the time of computation than both FSKSOR and HSKSOR point iterative methods. In the meantime, in terms of accuracy, the numerical solutions of 4HSEGKSOR method are in good agreement.

**Table 1** Numerical results of the iterative methods at $\alpha = 0.333, 0.666, 0.999$ for test problem 1

| M | | $\alpha = 0.333$ | | | $\alpha = 0.666$ | | | $\alpha = 0.999$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Method | k | t | Max error | k | t | Max error | k | t | Max error |
| 128 | FSKSOR | 404 | 3.20 | 2.5972e−02 | 283 | 3.11 | 1.3065e−02 | 166 | 3.14 | 1.2480e−03 |
| | HSKSOR | 199 | 1.88 | 2.5971e−02 | 140 | 1.87 | 1.3065e−02 | 86 | 1.87 | 1.2480e−03 |
| | 4HSEGKSOR | 102 | 1.84 | 2.5971E−02 | 71 | 1.82 | 1.3065E−02 | 43 | 1.82 | 1.2480E−03 |
| 256 | FSKSOR | 813 | 6.67 | 2.5972e−02 | 569 | 6.47 | 1.3065e−02 | 331 | 6.33 | 1.2480e−03 |
| | HSKSOR | 404 | 3.76 | 2.5972e−02 | 283 | 3.72 | 1.3065e−02 | 166 | 3.68 | 1.2480e−03 |
| | 4HSEGKSOR | 202 | 3.61 | 2.5972E−02 | 141 | 3.60 | 1.3065E−02 | 82 | 3.60 | 1.2480E−03 |
| 512 | FSKSOR | 1621 | 14.51 | 2.5972e−02 | 1136 | 13.63 | 1.3065e−02 | 659 | 13.08 | 1.2480e−03 |
| | HSKSOR | 813 | 7.71 | 2.5972e−02 | 569 | 7.59 | 1.3065e−02 | 331 | 7.45 | 1.2480e−03 |
| | 4HSEGKSOR | 403 | 7.32 | 2.5972E−02 | 277 | 7.27 | 1.3065E−02 | 163 | 7.22 | 1.2480E−03 |
| 1024 | FSKSOR | 3246 | 34.07 | 2.5972e−02 | 2269 | 30.86 | 1.3065e−02 | 1311 | 27.7 | 1.2480e−03 |
| | HSKSOR | 1621 | 16.27 | 2.5972e−02 | 1136 | 15.76 | 1.3065e−02 | 659 | 15.25 | 1.2480e−03 |
| | 4HSEGKSOR | 801 | 14.92 | 2.5972E−02 | 551 | 14.77 | 1.3065E−02 | 320 | 14.50 | 1.2480E−03 |
| 2048 | FSKSOR | 6365 | 87.08 | 2.5972e−02 | 4519 | 75.36 | 1.3065e−02 | 2618 | 63.69 | 1.2480e−03 |
| | HSKSOR | 3246 | 36.10 | 2.5972e−02 | 2269 | 33.96 | 1.3065e−02 | 1311 | 31.89 | 1.2480e−03 |
| | 4HSEGKSOR | 1593 | 31.32 | 2.5972E−02 | 1095 | 30.47 | 1.3065E−02 | 633 | 29.48 | 1.2480E−03 |

**Table 2** Numerical results of the iterative methods at $\alpha = 0.333, 0.666, 0.999$ for test problem 2

| M | | $\alpha = 0.333$ | | | $\alpha = 0.666$ | | | $\alpha = 0.999$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Method | k | t | Max error | k | t | Max error | k | t | Max error |
| 128 | FSKSOR | 385 | 3.23 | 3.4734e−04 | 264 | 3.13 | 5.2935e−04 | 144 | 3.09 | 4.5594e−04 |
| | HSKSOR | 193 | 1.85 | 9.3060e−04 | 135 | 1.84 | 1.1077e−03 | 84 | 1.84 | 1.0292e−03 |
| | 4HSEGKSOR | 94 | 1.80 | 9.3060E−04 | 67 | 1.80 | 1.1077E−03 | 39 | 1.81 | 1.0292E−03 |
| 256 | FSKSOR | 764 | 6.66 | 2.0160e−04 | 513 | 6.40 | 3.8483e−04 | 291 | 6.20 | 3.1267e−04 |
| | HSKSOR | 385 | 3.80 | 3.4734e−04 | 274 | 3.77 | 5.2935e−04 | 144 | 3.73 | 4.5594e−04 |
| | 4HSEGKSOR | 183 | 3.71 | 3.4734E−04 | 126 | 3.67 | 5.2936E−04 | 73 | 3.65 | 4.5594E−04 |
| 512 | FSKSOR | 1467 | 13.99 | 1.6516e−04 | 1025 | 13.61 | 3.4871e−04 | 576 | 12.74 | 2.7687e−04 |
| | HSKSOR | 764 | 7.65 | 2.0159e−04 | 513 | 7.56 | 3.8483e−04 | 291 | 7.40 | 3.1267e−04 |
| | 4HSEGKSOR | 360 | 7.32 | 2.0159E−04 | 247 | 7.21 | 3.8484E−04 | 141 | 7.18 | 3.1268E−04 |
| 1024 | FSKSOR | 2906 | 31.95 | 1.5605e−04 | 2049 | 30.49 | 3.3969e−04 | 1125 | 27.04 | 2.6791e−04 |
| | HSKSOR | 1467 | 16.23 | 1.6516e−04 | 1115 | 15.91 | 3.4871e−04 | 654 | 15.44 | 2.7687e−04 |
| | 4HSEGKSOR | 684 | 15.12 | 1.6516E−04 | 478 | 14.90 | 3.4871E−04 | 255 | 14.69 | 2.7686E−04 |
| 2048 | FSKSOR | 5708 | 80.73 | 1.5380e−04 | 4056 | 71.23 | 3.3742e−04 | 2155 | 60.45 | 2.6568e−04 |
| | HSKSOR | 2905 | 35.44 | 1.5605e−04 | 2049 | 33.81 | 3.3969e−04 | 1151 | 31.90 | 2.6792e−04 |
| | 4HSEGKSOR | 684 | 29.65 | 1.6516E−04 | 478 | 29.33 | 3.4871E−04 | 255 | 29.20 | 2.7686E−04 |

**Table 3** Numerical results of the iterative methods at $\alpha = 0.333, 0.666, 0.999$ for test problem 3

| M | | $\alpha = 0.333$ | | | $\alpha = 0.666$ | | | $\alpha = 0.999$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Method | k | t | Max error | k | t | Max error | k | t | Max error |
| 128 | FSKSOR | 423 | 3.16 | 1.1757e−03 | 295 | 3.13 | 2.5780e−03 | 172 | 3.11 | 2.2070e−03 |
| | HSKSOR | 209 | 1.85 | 1.1785e−03 | 145 | 1.84 | 2.5807e−03 | 89 | 1.84 | 2.2096e−03 |
| | 4HSEGKSOR | 104 | 1.83 | 1.1785E−03 | 73 | 1.83 | 2.5807E−03 | 43 | 1.83 | 2.2096E−03 |
| 256 | FSKSOR | 853 | 6.69 | 1.1750e−03 | 592 | 6.44 | 2.5773e−03 | 342 | 6.27 | 2.2064e−03 |
| | HSKSOR | 423 | 3.82 | 1.1757e−03 | 295 | 3.78 | 2.5780e−03 | 172 | 3.75 | 2.2071e−03 |
| | 4HSEGKSOR | 207 | 3.68 | 1.1757E−03 | 145 | 3.68 | 2.5780E−03 | 84 | 3.65 | 2.2071E−03 |
| 512 | FSKSOR | 1707 | 14.43 | 1.1748e−03 | 1181 | 13.68 | 2.5772e−03 | 681 | 12.89 | 2.2063e−03 |
| | HSKSOR | 853 | 7.70 | 1.1750e−03 | 592 | 7.59 | 2.5773e−03 | 342 | 7.43 | 2.2064e−03 |
| | 4HSEGKSOR | 412 | 7.31 | 1.1750E−03 | 287 | 7.24 | 2.5773E−03 | 167 | 7.20 | 2.2064E−03 |
| 1024 | FSKSOR | 3392 | 33.8 | 1.1748e−03 | 2353 | 30.74 | 2.5771e−03 | 1356 | 27.92 | 2.2062e−03 |
| | HSKSOR | 1707 | 16.51 | 1.1748e−03 | 1181 | 16.02 | 2.5772e−03 | 681 | 15.46 | 2.2063e−03 |
| | 4HSEGKSOR | 820 | 15.16 | 1.1748E−03 | 562 | 14.96 | 2.5772E−03 | 330 | 14.78 | 2.2063E−03 |
| 2048 | FSKSOR | 6774 | 88.57 | 1.1747e−03 | 4680 | 75.59 | 2.5771e−03 | 2708 | 63.53 | 2.2062e−03 |
| | HSKSOR | 3392 | 36.43 | 1.1748e−03 | 2353 | 34.29 | 2.5771e−03 | 1356 | 32.55 | 2.2062e−03 |
| | 4HSEGKSOR | 820 | 29.58 | 1.1748E−03 | 562 | 29.40 | 2.5772E−03 | 330 | 29.23 | 2.2063E−03 |

**Table 4** Number of iterations (Iter) and computational time depreciation rate of the HSKSOR versus FSKSOR iterative methods

| Example | | $\alpha = 0.333$ | $\alpha = 0.666$ | $\alpha = 0.999$ |
|---|---|---|---|---|
| 1 | Iter | 49.00–50.74% | 49.79–50.53% | 48.19–49.92% |
| | Time | 41.25–58.54% | 39.87–54.94% | 40.45–49.93% |
| 2 | Iter | 47.92–49.87% | 46.59–49.48% | 41.67–50.52% |
| | Time | 42.72–56.10% | 41.09–52.53% | 39.84–47.23% |
| 3 | Iter | 49.68–50.59% | 49.72–50.85% | 48.26–49.93% |
| | Time | 41.46–58.87% | 41.21–54.64% | 40.19–48.76% |

**Table 5** Number of iterations (Iter) and computational time depreciation rate of the 4HSEGKSOR versus FSKSOR iterative methods

| Example | | $\alpha = 0.333$ | $\alpha = 0.666$ | $\alpha = 0.999$ |
|---|---|---|---|---|
| 1 | Iter | 74.75–75.32% | 74.91–75.77% | 74.10–75.82% |
| | Time | 42.50–64.03% | 41.48–59.57% | 42.04–53.71% |
| 2 | Iter | 75.46–88.02% | 74.62–88.21% | 72.92–88.17% |
| | Time | 44.27–63.27% | 42.49–58.82% | 41.13–51.70% |
| 3 | Iter | 75.41–87.89% | 75.25–87.99% | 75.00–87.81% |
| | Time | 42.09–66.60% | 41.53–61.11% | 41.16–53.99% |

# References


1. Ionescu C, Lopes A, Copot D, Machado JAT, Bates JHT (2017) The role of fractional calculus in modeling biological phenomena: a review. Commun Nonlinear Sci Numer Simul 51:141–159
2. Kumar D, Rai KN (2017) Numerical simulation of time fractional dual-phase-lag model of heat transfer within skin tissue during thermal therapy. J Therm Biol 67:49–58
3. Khan NA, Razzaq OA, Mondal SP, Rubbab Q (2019) Fractional order ecological system for complexities of interacting species with harvesting threshold in imprecise environment. Adv Differ Equ 2019(1):405
4. Nadzharyan TA, Sorokin VV, Stepanov GV, Bogolyubov AN, Kramarenko EYu (2016) A fractional calculus approach to modeling rheological behavior of soft magnetic elastomers. Polymer 92:179–188
5. Luc NH, Huynh LN, Tuan NH (2019) On a backward problem for inhomogeneous time-fractional diffusion equations. Comput Math Appl 78(5):1317–1333
6. Tuan NH, Ngoc TB, Tatar S (2018) Recovery of the solute concentration and dispersion flux in an inhomogeneous time fractional diffusion equation. J Comput Appl Math 342:96–118
7. Zheng G-H (2014) Recover the solute concentration from source measurement and boundary data. Inverse Probl Sci Eng 23(7):1199–1221
8. Tasbozan O, Esen A, Yagmurlu NM, Ucar Y (2013) A numerical solution to fractional diffusion equation for force-free case. Abstract Appl Anal, Article ID 187383
9. Badr M, Yazdani A, Jafari H (2018) Stability of a finite volume element method for the time-fractional advection-diffusion equation. Numer Methods Partial Differ Equ 34(5):1459–1471
10. Liu F, Zhuang P, Turner I, Burrage K, Anh V (2014) A new fractional finite volume method for solving the fractional diffusion equation. Appl Math Model 38:3871–3878
11. Saw V, Kumar S (2020) Collocation method for time fractional diffusion equation based on the Chebyshev polynomials of second kind. Int J Appl Comput Math 6(4):117

12. Akram T, Abbas M, Ismail AI (2019) An extended cubic B-spline collocation scheme for time fractional sub-diffusion equation. AIP Conf Proc 2184:060017
13. Yu H, Wu B, Zhang D (2019) The Laguerre-Hermite spectral methods for the time-fractional sub-diffusion equations on unbounded domains. Numer Algorithms 82(4):1221–1250
14. Guo C, Zhao F (2019) Numerical methods for the time fractional diffusion equation. J Phys Conf Ser 1324(1):012014
15. Hong EJ, Saudi A, Sulaiman J (2017) Numerical analysis of the explicit group iterative method for solving poisson image blending problem. Int J Imaging Robot 17(4):15–24
16. Saudi A, Sulaiman J (2016) Path planning simulation using harmonic potential fields through four point-EDGSOR method via 9-point Laplacian. J Teknol 78(8–2):12–24
17. Dahalan AA, Muthuvalu MS, Sulaiman J (2013) Numerical solutions of two-point fuzzy boundary value problem using half-sweep alternating group explicit method. AIP Conf Proc 1557:103–107
18. Akhir MKM, Othman M, Sulaiman J, Majid ZA, Suleiman M (2011) Numerical solution of Helmholtz equation using a new four-point EGMSOR iterative method. Appl Math Sci 5(77–80):3991–4004
19. Rahman R, Mat Ali NA, Sulaiman J, Muhiddin FA (2019) Block iterative method for the solution of fractional two-point boundary value problems. J Phys Conf Ser 1358(1):012053
20. Rahman R, Ali NAM, Sulaiman J, Muhiddin FA (2019) Application of the half-sweep EGSOR iteration for two-point boundary value problems of fractional order. Adv Sci Technol Eng Syst 4(2):237–243
21. Youssef IK (2012) On the successive over relaxation method. J Math Stat 8(2):176–184
22. Suardi MN, Radzuan N, Sulaiman J (2017) Cubic B-spline solution of two-point boundary value problem using HSKSOR iteration. Glob J Pure Appl Math 13(11):7921–7934
23. Radzuan NZFM, Suardi MN, Sulaiman J (2017) KSOR iterative method with quadrature scheme for solving system of Fredholm integral equations of second kind. J Fundam Appl Sci 9(5S)
24. Podlubny I (1999) Fractional differential equations. Academic Press, Cambridge
25. Zahra WK, Elkholy SM (2013) Cubic spline solution of fractional Bagley-Torvik equation. Electron J Math Anal Appl 1(2):230–241
26. Young DM (1971) Iterative solution of large linear systems. Academic Press, New York
27. Hadjidimos A (2000) Successive over relaxation (SOR) and related methods. J Comput Appl Math 123(1–2):177–199
28. Muhiddin FA, Sulaiman J, Sunarto A (2019) MKSOR iterative method for the Grünwald implicit finite difference solution of one-dimensional time-fractional parabolic equations. AIP Conf Proc 2138:030026
29. Uddin M, Haq S (2011) RBF's approximation method for time fractional partial differential equations. Commun Nonlinear Sci Numer Simul 16(11):4208–4214
30. Jiang Y, Ma J (2011) High-order finite element methods for time-fractional partial differential equations. J Comput Appl Math 235(11):3285–3290
31. Ma Y (2014) Two implicit finite difference method for time fractional diffusion equation with source term. J Appl Math Bioinf 4(2):125–145

# Fake News Detection

**Si Hong Long and Mohd Pouzi Bin Hamzah**

**Abstract** Everyday people receive a lot of information through social media and online news portals. To distinguish whether the information is fake or true is a big problem. An algorithm has been developed to distinguish fake news and true news by searching the relevant news from reliable news website based on the news given. This results in the similarity percentage between news and the relevant news. The algorithm has been tested with the dataset collected by Dr. Victoria L. Rubin that consists of 180 true news and 180 fake news from several American and Canadian news websites. The precision of 69.44% has been achieved with the dataset.

**Keywords** Fake news · Cosine similarity · Natural language processing

## 1 Introduction

The meaning of fake news is false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke. Nowadays the amount of fake news keeps increasing especially news regarding Covid-19, but people cannot distinguish between the true and fake news. In large part, a deeper concern that the prevalence of "fake news" has increased political polarization, decreased trust in public institutions, and undermined democracy [1]. An example of fake news such as in 2016, a Facebook post about nationwide order banning The Pledge of Allegiance in schools in the United States that had been signed by President Obama was shared and commented upon a total of 2.2 million times on Facebook [2]. There are two examples of fake news that happened in Malaysia that have been reported in New Straits Times. First, the news regarding PT3 papers for

S. H. Long · M. P. B. Hamzah (✉)
Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
e-mail: mph@umt.edu.my

S. H. Long
e-mail: ahonglong960806@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_25

subject English, Mathematics, Science and Integrated Living Skills reported leaked [3]. Second, the news regarding October beer festival in Terengganu [4].

The type of fake news can be divided into four types which are actual "fake news", satire news, poorly reported news and misleading news. The actual "fake news" are stories that are completely made up and do not happen in the world. Satire news are fake articles that are meant for humour and without prove. Poorly reported news is news that are reported badly but not completely made up. The misleading news are news that try to change the perspective of readers toward a topic [5].

People are receiving information every day, but they do not have the ability to recognize the fake news. An exclusive Ipsos poll conducted for BuzzFeed News found that 75% of American adults who always use Facebook as the source of news are likely to believe fake news headline than those who do not use Facebook as source for news [6]. According to estimate made by police and by local community leaders, 86 to 238 Berom ethnic minority were killed in Gashish between 22 and 24 June 2018, just because a fake Facebook news posted by a man in United Kingdom that Fulani Muslims were killing Christians [7].

In order to overcome the problem stated above, there is a need to develop an algorithm that can help people to distinguish fake news and true news by comparing the news from user with several reliable news website and give the related news from reliable news website as references to the user.

## 2 Related Work

Sirajudeen et al. [8] proposed three-phase method to detect online fake news that use java programming language. The first phase is checking IP validity. The second phase is checking the content of the information of online news such as article, title, author and background information of the article with a database that contain the verification information. The third phase is to determine the status of fake news based on result from two previous phases. Another approach proposed by Gahirwal et al. [9] is by comparing the headlines and compare news article with top search. Feyza and Bilal [10] proposed a two-step method for identifying fake news in social media and the method was tested on three real data sets in terms of different evaluation metrics.

Granik and Mesyura [1] proposed an algorithm making use of naïve Bayes classifier. This approach achieved accuracy approximately 74% on test set. They found that spam messages and fake news article have common properties like a lot of grammatical mistakes, emotionally coloured, often use same set of word and affect reader's opinion on some topic in manipulative way. The main idea is to treat each word of the news article independently. Wei and Wan [11] introduced a method that uses class sequential rules (CSR) and basic features (body-independent features) extracted from headline to train support vector machine (SVM) classifier. They also add body dependent features such as Informality, Sentiment, InformalGap, sentiGap,

Similarity, Recognizing Textual Entailment (RTE) to train SVM. The SVM toolkit is from the scikit-learn.

Ahmed [12] introduced feature extraction using term frequency (TF) and term frequency-inverted document frequency (TF-IDF). Other features are keystroke such as editing patterns and timespan, n-grams features and semantic similarity. The experiments involve six different machine learning algorithms which are Stochastic Gradient Descent (SGD), K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machines (SVM) and Linear Support Vector Machines (LSVM). The experiment also studies the impact of n-grams size on performance. Total of four experiment were carried out. From the first experiment, Linear-based classifiers such as Linear SVM, Logistic regression and SDG yield better result than nonlinear methods. The accuracy increases as number of feature values increase. As the size of n-grams increases, the accuracy will decrease. The performance for TF-IDF is better than TF. KNN achieve lowest accuracy which is 47.2% with 4-g word and 50,000 feature values. From the second experiment, they found that Linear-based classifier is still better than nonlinear as Linear SVM achieved accuracy as 92%. The performance of Linear SVM is not affected by number of feature values, but as size of n-gram increase the accuracy decrease. From the third experiment, they found that keystrokes feature with n-gram yielded better accuracy. From the fourth experiment, they found that as the percentage of change increase the semantic measurement decrease.

Most of the researchers use artificial intelligence in online fake news detection. In our work, we propose a new approach by comparing news article from the user and the articles from reliable news sources.

## 3 Methodology

Figure 1 shows the algorithm of fake news detection system. The headline, article or URL are provided by the user as an input to the system. Then the article and the headline will be extracted from the webpage. Next, all related article will be retrieved from the reliable news website based on the headline. Then data pre-processing such as lemmatization, stop word removal will be carried out. After the pre-processing, the algorithm proceeds with the calculation of the Term Frequency-Inverted Document Frequency (TF-IDF) and cosine similarity between news. Finally, percentage of similarity will be displayed.

### *3.1 Lemmatization*

Lemmatization is to improve the performance of natural language processing by generating the word into its root word. For example, playing, plays and played after

1) Get article, headline or URL from user.
2) If user use headline and article to check news, then use headline as keywords to find similar news.
3) If user use article only to check news, if the length of article longer than 20 word, use first 20 word as keywords to find similar news.
4) If user use URL to check news, extract headline and article from given URL, use headline as keywords to find similar news.
5) Remove all the punctuation from keywords and declare article as String.
6) Declare three list all_news_article, reference_list and all_news_url.
7) Pre-process article and add to all_news-article list.
8) Replace all the space in keywords with symbol "+".
9) Find all the URL link that relate with keyword from reliable news website with web scraping and add into all_news_URL list.
10) Extract all article and headline from all the URL in all_news_URL list, then add pre-process article into all_news_article, and add article, headline and URL into reference_list.
11) Calculate Term Frequency-Inverted Document Frequency.
12) Calculate cosine similarity.
13) Convert cosine similarity into percentage with 2 decimal places.
14) Add similarity to reference_list and sort list in descending order base on similarity.
15) Find the similarity of first news in reference list if more than or equal to 70, then delete the rest of news in reference_list.
16) If first news in reference_list not equal to 0, check the news with 0 percent similarity and delete it.
17) Calculate average similarity from reference_list.
18) If average similarity greater or equal to 70, then status is "The news is true".
19) If average similarity smaller then 70, then status is "The news is not reliable".
20) Display the average similarity, status and news in reference_list to user.

**Fig. 1** Fake news detection system algorithm

lemmatization will become play. All the lemmas are the English words that can be found in dictionaries.

## *3.2 Stop Words and Punctuation Removal*

Stop words are the words that do not provide important information to document and common to most documents. Stop words will decrease the performance in natural language processing. The example of stop words are 'as', 'the', 'be', 'are' and etc.

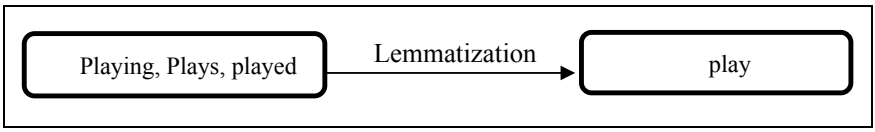



**Fig. 2** Example of lemmatization

Like the stop words, the punctuation is not important in natural language processing and it will affect the performance. The example of punctuations are (?, !, ;, :, ‘, “).

### *3.3 Term Frequency-Inverted Document Frequency (TF-IDF)*

In information retrieval, TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Fig. 2).

In mathematical representation, TF-IDF = TF * IDF which is term frequency multiply with inverted document frequency. The formula for Term frequency is the number of a word appear in a document divide the total number of words in document. The formula for Inverse document frequency is the total number of documents in corpus by document frequency for each term and apply logarithmic scaling on the result.

$$\mathrm{TF} - \mathrm{IDF} = \frac{A}{B} \times \left(1 + \log \frac{C}{1+D}\right) \tag{1}$$

where A represents the number of a word appear in a document, B represents the total number of words in a document, C represents total number of documents in corpus and D represents number of documents a word appears.

### *3.4 Cosine Similarity*

Cosine similarity is used to measure the cosine angle between two terms as they represented in their vectorized forms and non-zero positive vectors in an inner product space. The term vectors are close to each other and in the same direction, thus the score is closer to 1 (cos 0°), mean they are similar. If the term vectors score close to 0 (cos 90°), mean they are not similar. Term vectors score close to −1 (cos 180°), mean they are unrelated, and they are in opposite direction to each other.

Cosine similarity are dot product of the two term vectors $u$ and $v$, divided by the product of their L2 norms. The mathematically representation of dot product between two vectors as shown in Fig. 3.

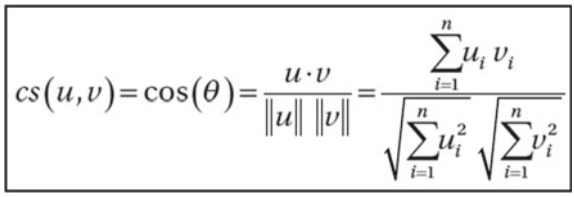



**Fig. 3** The formula for cosine similarity. *Source* [13]

$$u \cdot v = |u||v| \cos(0) \tag{2}$$

The cosine similarity can be derived from the above formula where $u_i$ represents the various features of term vector $u$, $v_i$ represents the various features of term vector $v$ and n represents the total number of features.

## 3.5 Dataset

The dataset used in this research was collected by the Language and Information Technology Research Lab directed by Dr. Vitoria Rubin, Western University, London, Ontario, Canada [14]. The dataset consists of 360 news from several United State and Canada news websites. The dataset is divided into two sets.

Set 1 consists of 240 articles in which 120 out of 240 articles are legitimate, and the remaining articles are satirical. The legitimate news set is collected from The New York Times of United State and The Toronto Star of Canada. The satirical news set is collected from The Onion of United State and The Beaverton of Canada.

Set 2 consists 120 articles, half of it is legitimate and the others are satirical. The legitimate news set is collected from USA Today, The Wall Street Journal, The Los Angeles Times, The New York Post, Newsday, The Denver Post of United State and The Globe and Mail, The Vancouver Sun, The Calgary Herald, The National Post, The Edmonton Journal, The Hamilton Spectator of Canada. The satirical news set is collected from The Daily Currant, The Spoof, The National Report, The People's Cube, World News Daily Report, Urban Anomie of United State and CBC's Punchline, The Codfish, The Lapine, The Syrup Trap, Sage News of Canada.

## 3.6 Evaluation

To evaluate the performance, precision has been used as a measurement. The calculation of precision is using #(*true_positive*) divided by #(*true_positive*,*false_positive*). #(*true_positive*) is the number of news correctly classified by the approach. #(*true_positive*,*false_positive*) is the total number of news in the dataset. The formula is as shown below

$$Precision = \frac{\#(true\ positive)}{\#(true\ positive, false\ positive)} \quad (3)$$

## 4 Results and Discussion

Dataset consists of 360 news; 180 news are legitimate (true news) and 180 news are satirical (fake news). After testing with the dataset, the algorithm is able to classify correctly all the fake news but can only classify correctly 70 out of 180 for true news. An average precision of 69.44% is achieved with the dataset.

## 5 Prototype

The interfaces of prototype are as shown in Figs. 4 and 5. Figure 4 is the input interface for the prototype. There are two section and one button in this interface which are checking news by using and related information about the news. The user can choose either by using URL only, article only or both headline and article for checking the news.

The result of news checking is as shown in Fig. 5. The result interface will show six information. The similarity between the news provided by user with the news from pre-defined reliable news sources. The comment either the news is fake or true. The headline, article, URL and similarity of news from reliable news sources

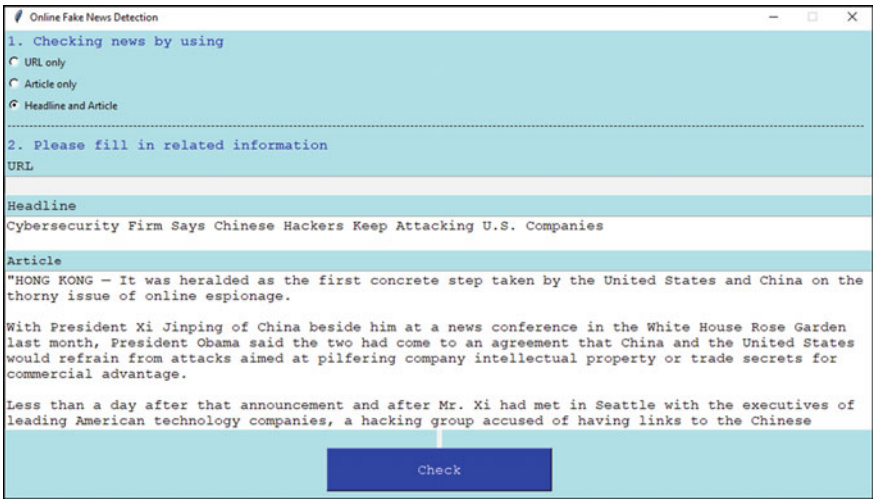



**Fig. 4** The input interface

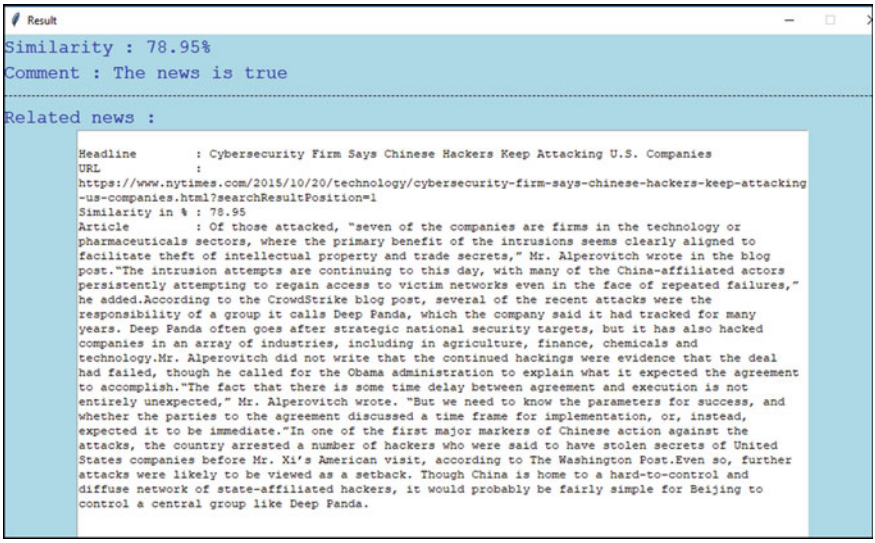



**Fig. 5** Result interface

related to user news if available. If no related news from reliable news sources, then the prototype will show an alert box with content 'Sorry that we cannot find news related to your topic. Your news has high probability of FAKE NEWS'.

## 6 Conclusion

The fake news tries to change the reader perspective toward a specific topic. We had proposed an algorithm to detect fake news by comparing headline and article with several reliable news website. Our algorithm can achieve a precision of 69.44%. There are three suggestions to improve the algorithm. First, is to increase number of reliable news website, so that comparison can be made with higher accuracy. Second, the algorithm can detect the place and from the place name fake news detection can be localized. This can limit the news sources and speed up the time needed for checking. For example, if Kuala Lumpur is stated in the news article, the algorithm will identify the country as Malaysia and will then retrieve the reliable news from websites in Malaysia such as The Star Online to classify the news provided by the user.

## References


1. Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In: Proceedings of IEEE 1st Ukraine conference on electrical and computer engineering, UKRCON 2017, pp 900–903
2. Roberts H (2016) This is what fake news actually looks like—we ranked 11 election stories that went viral on Facebook. Retrieved from https://www.businessinsider.com/fake-presidential-election-news-viral-facebook-trump-clinton201611/?IR=T
3. Abas A (2018) PT3 papers not leaked, rumours untrue. New Straits Times, 9 Oct 2018. Retrieved from https://www.nst.com.my/news/nation/2018/10/419352/pt3-papers-not-leaked-rumours-untrue
4. Nizam F (2018) No Oktoberfest in Terengganu. New Straits Times, 12 Oct 2018. Retrieved from https://www.nst.com.my/news/nation/2018/10/420604/no-oktoberfest-terengganu
5. Ashe S (2017) The 4 types of 'Fake News'. Observer, 1 Apr 2017. Retrieved from https://observer.com/2017/01/fake-news-russia-hacking-clinton-loss/
6. Silverman C, Singer-Vine J (2016) Most Americans who see fake news believe it, new survey says. Retrieved from https://www.buzzfeednews.com/article/craigsilverman/fake-news-survey
7. Adegoke Y, BBC Africa Eye (2018). Nigerian say "fake news" on Facebook is killing people. Retrieved from https://www.bbc.co.uk/news/resources/idtsh/nigeria_fake_news
8. Sirajudeen SM, Azmi NFA, Abubakar AI (2017) Online fake news detection algorithm. J Theor Appl Inf Technol 95(17):4114–4122
9. Gahirwal M, Moghe S, Kulkarni T, Khakhar D, Bhatia J (2018) Fake news detection. Int J Adv Res Ideas Innov Technol 4(1):817–819
10. Feyza AO, Bilal A (2020) Fake news detection within online social media using supervised artificial intelligence algorithms. Phys A Stat Mech Appl 540. https://doi.org/10.1016/j.physa.2019.123174
11. Wei W, Wan X (2017) Learning to identify ambiguous and misleading news headlines. In: IJCAI International joint conference on artificial intelligence, pp 4172–4178. https://doi.org/10.24963/ijcai.2017/583
12. Ahmed H (2017) Detecting opinion spam and fake news using N-gram analysis and semantic similarity by. MSc thesis, University of Victoria
13. Sarkar D (2016) Text analytics with Python: a practical real-world approach to gaining actionable insight from your data. New York
14. Rubin VR, Conroy NJ, Chen Y, Cornwell S (2016) Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of the workshop on computational approaches to deception detection at the 15th annual conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-CADD2016), San Diego, California, 17 June 2016

# A Literature Review on Text Classification and Sentiment Analysis Approaches


**Wang Dawei, Rayner Alfred, Joe Henry Obit, and Chin Kim On**

**Abstract** Sentiment analysis is an important branch task of text classification and the related system usually is applied to in perception of user emotion and public opinion monitoring. By comparison, the text classification can be applied to more fields than sentiment analysis. In the system architecture, same as text classification, the complete classification system mainly contains data acquisition, data pre-process, feature extraction, classification algorithm and result output. The Web crawler usually be used in first step, the URL Link, hashtags, Non-Chinese text should be removed in second step. In feature extraction, the IG, TF-IDF, Word2vec usually be used. Then, the SVM, Naive Bayes, KNN or Neural network algorithm usually be used in classifier. Furthermore, as a system that can run automatically, the sentiment analysis system should be able to extract significant feature from corpus and make accurately analysis about emotional polarity of text corpus. At present, the system improvement direction of related system focuses on 3 aspects: data acquisition, feature extraction and classifier algorithm.

**Keywords** Literature review · Text classification · Sentiment analysis · Classifier algorithm

W. Dawei
Hebei University of Engineering, Handan, China
e-mail: wangdawei853@gmail.com

R. Alfred (✉) · J. H. Obit · C. K. On
Knowledge Technology Research Unit, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88000 Kota Kinabalu, Sabah, Malaysia
e-mail: ralfred@ums.edu.my

J. H. Obit
e-mail: joehenry@ums.edu.my

C. K. On
e-mail: kimonchin@ums.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_26

## 1 Introduction

In information era, the development of Web2.0 is very speed, it encourage people issue their content on the social media and comment to other people's content, so, users in social media not only search information, but also generate novel content [1]. In China, Sina Weibo is a one of famous online social network. It incorporated the benefits of multiple social media such as blogs, traditional website, and online forum. According to official statistic, in 2018 years, nearly 70 million people each month become active user on Weibo and the total user on Weibo has reach to 462 million in December 2019 already. In addition, it is should be noted that daily active user of Weibo is 200 million in 2019 and the number is 28 million more than 2017 [2].

In aboard, as the number of social media site increased, for example, Twitter and Instagram have more meanings instead of only a website, which save user's content [3]. It shows that the social media is more and more important. Just as X. Yang said: in information ear, twitter has a great ability to feel people's emotion at any time through huge UGC (user generate content) and even the enterprise development, government can be affected by Twitter [4]. Hence, more tools were generated for making further data mining and analysis about the Web text. Among them, the text classification system is a kind of tool that can make full use of the user's content. Meanwhile, text classification can make people liberate from heavy work of manually [5]. Sentiment analysis was also introduced in health care using natural language processing (NLP) [6]. In essence, sentiment analysis is also text classification, but its mission objective is different from text classification. There are both similarities and differences between the two systems. Detailed understanding these can better to help us carry out the next step of research work. So, the following research questions will be focused in the paper:

1. RQ1: In application area, what actually each part in text classification system and sentiment analysis system?
2. RQ2: What is the proposal in next step of research about related system?

In general, a typical text classification system always contains data input, data pre-process, feature extraction, data classification and result output. Thus, the performance of the text classification can be improved by improving the processes involved in these five modules. In this paper, a comprehensive review about text classification system will be reviewed. It is hoped that this review can assist researcher to understand about the current text classification system.

## 2 Literature Search

In information era, the social media has more and more important and has produced significant influence on many aspects of human society. Just as [7] said: in last ten years, social media has become an important role for accessing and transmit

information in many fields, business for example, entertainment, science and crisis management and so on [8]. Thus, performing social media data analysis has higher value than before.

For the types of data, social media platform allows user to post variety data types such as text, picture, video, sound and geolocation [9]. Most of the social media data are in the form of texts format [3]. Thus, the text data on social media still has huge influences on our daily decision making [10].

## *2.1 Database Selection*

China National Knowledge Infrastructure (CNKI) is the largest database of Chinese academic articles [11]. IEEE Xplore database mainly includes engineering articles and it certainly is the main search platform of text analysis system. In addition, the Science Direct database is an important database also. The databases and fields considered are listed in Table 1.

Firstly, the "text classification system" was used as the search keyword in CNKI and the returned result was 715 articles. Then, all of papers are then filtered according to the relevancy of the articles. It should be noted that some English journals were input in CNKI recently years. So, 249 English papers were in the original result. Through checking, in Chinese article, some articles that focus on Tibetan, Uighur were remove and some article that unrelated to search keyword were remove. There are 401 articles were kept. In English article, there are 52 articles were kept. Secondly, the "sentiment analysis system" was used as the search keyword in CNKI, the field is theme also. The returned result is 286. There are 190 articles were kept after checking.

In IEEE Xplore, the "text classification system" was used as keyword in search. The returned result in journal and magazine is 434. The paper thinks that the number of articles is enough to show the research situation about text classification system. There are 370 articles were kept after checking. Then, the "sentiment analysis system" was used as key word in search in IEEE Xplore. The returned result in journal and magazine is 218. There are 133 articles were kept after checking.

**Table 1** Keywords and databases which were used for the review

| Search terms | | | Databases | Fields |
|---|---|---|---|---|
| "Sentiment analysis" | AND | System | CNKI | Theme |
| OR "text analysis" | | | IEEE Xplore | |
| OR "text classification"<br>OR "sentiment classification" | | | | |

# 3 Comparative Study

## *3.1 Papers*

Take the number of related issued paper in CNKI as example. There are 453 articles are related text classification system and only 190 articles are related to sentiment analysis system. It shows that in spite of both belong to text classification, the researcher of text classification system is much more than sentiment analysis system. Furthermore, the situation reveals that the application field of text classification system is more extensive than sentiment analysis.

## *3.2 Text Classification and Sentiment Analysis*

As one of important web tool, text classification system was generated and been widely application. Furthermore, the system includes the follow characteristics:

**Automation Running** Text classification system usually is used in public opinion monitor, E-commerce and news classification. The related text was obtained from web. Before classification, the text data will be through preprocessing, word segmentation, then, processed data will be make classification and output [12]. Thus, firstly, the system is automated, and it can get data automatically. Then the text processing, feature extraction and result output can be automated running also.

**Range of Applications** A classification system for quality assessment has been designed to handle ambiguities of texts [13]. A text analysis system has also been designed to detect cyber terrorism based on the contents of the web documents containing the information related to cyber terrorism [14].

A text classification system has also been designed and implemented to analyze social media contents to detect earthquake related text [15]. The basic flow is that user may download the reports of earthquake area, assessment report and background report from official website. Then, the system classifies the documents into three types of documents based on keyword classification of Boolean model.

A text classification system has also been designed to analyze literature [16]. The article proposed a solution that combines library and classifier. The "Chinese Library Classification" was treated as standard of classifier parameter in the article for extracting subject related word and the kernel function was used in $k$-NN for solving the overlap boundary problem, classifier's precise decision problem. So, the engineering and social science is well integrated in the article. Therefore, the text classification system has been application in many fields.

In [17], according to the investigation of professional organizations at home and abroad in recent years, it is shown that the reviews provided by the common online shopping users play an important role in the purchase decision of potential consumers.

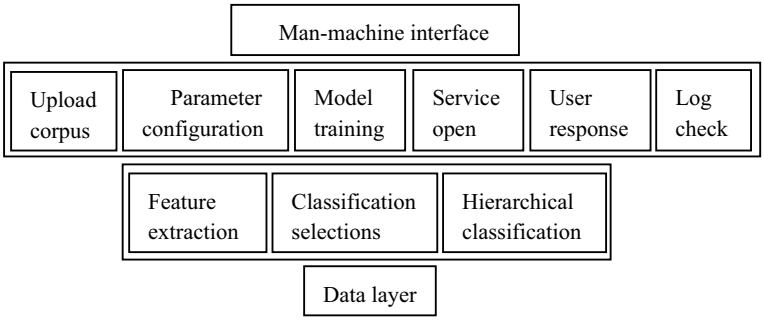



**Fig. 1** The structure of text classification system [21]

These reviews are important basis for the potential consumers to decide on whether to buy or not. So, a deep learning model was proposed in the article to perform the classification of emotional polarity for improving user experience.

In [18], a sentiment analysis system was used to monitor the public opinion in industries and provide guidance channel to help people who in need of the industry. In [19], the WeChat was as research object. In the article, a novel intelligent customer service system is designed and developed based on sentiment analysis technology. The dialogue (text) in WeChat between system and customer can be acquired and analysis in real time and user sentiment can be perceived, analyzed.

**Structure of Text Classification System** A classification module has been developed to process massive short text [20]. There are two main interfaces included which are system administrator and user interfaces. The administrator of the system can access system, upload corpus, configure parameters of classifier, model training and service generate. User of the system can input short text and can get answer from the system. In addition to that, other modules in the system contain text representation and text classification modules. The basic structure can be visualized as shown in Fig. 1.

Figure 1 illustrates a typical structure in the classification system, and it can be designed and implemented to have a man-machine dialogue with user for text classification. A text classification system has been proposed that can be applied to analyze users' comments. Some components of the main system include the front end, classifier and corpus database. The training data used in the system is obtained from the expert classification systems and Term Frequency—Inverse Document Frequency (TF-IDF), *k*-Nearest Neighbors (*k*-NN) and language model (LM) were used to perform the classification task [22]. Therefore, a basic text classification system should always contain text acquisition, classifier and panel to show the results. In addition to that, administer should be able to adjust the values of parameters of the classifiers in order to get better results.

**Sentiment Analysis System** Sentiment analysis can be used to gauge the polarities of some products or issues [23–25]. A sentiment analysis system can be designed

that contains a web crawler, a data pre-process, a classification module and finally produces the result of the text classification. The main objective is to discover the sentiment polarity about user comments about certain products or concepts and these comments are mainly obtained from business web page [26]. This sentiment analysis system was designed not limited to sentiment analysis. It was a complete structure for operation and implementation of the entire system. The web crawler was used to download text data from special Web page to data warehouse and the Hadoop and Spark were used to do the big data analysis of the crawled text documents.

A unique sentiment classification model was also designed to analyze real-time public sentiments obtained from the Twitter related to the 2012 U.S. presidential candidates [27]. Twitter comments usually full of emotional text, so, they are very suitable to be used as resources for investigating the performance of the sentiment analysis tools. There are 4 basic modules in the system also, they are data acquisition, data process, data classification and result display. But there is some adjustment in the system architecture. Firstly, the style of result display is aggregate by candidate. Actually, it can be decided by the need of system design. Secondly, a baseline sentiment model was create by the system and used Amazon Mechanical Turk (AMT) to get as varied a population of annotators as possible [27].

In the above articles, the complete system basically has automatic function of data acquisition and processing. It has similar functions as the text classification system. But the application field of these two kinds of system is different. The application filed of text classification system is more widely than sentiment analysis.

## 4 Data Acquisition Module

Internet is the main source of text data and web spider or other grab tools can be used to crawl and obtain its content [28]. Web crawlers are usually used in the data acquisition process and it is the first step in sentiment analysis system. Crawler is a kind of automatic program and the general structure of crawler is shown in Fig. 2.

The general web crawler accesses webpage according to predefined seed URL and download customized content. Then, the page parsing will remove the html tag and

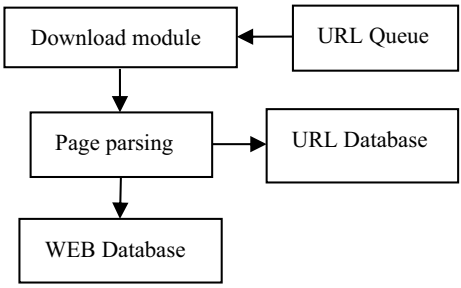



**Fig. 2** Architecture of general web crawler [29]

keep the abstract, URL in Web database. Meanwhile, the next URL will be extract until reach the certain stop conditions of the system. In a word, for sentiment analysis system, the process of accessing page is that traversing text information from social media.

A practical distributed crawler has the following characteristics [30]: (1) a complete code of web crawler should be used in many web page and many repetitive work must be contained in each web crawler for increasing efficiency and easy to management. (2) Easy to be extend. For each of new target website, the number of codes in web crawler is not many and do not need to change the level of source code. (3) Fault tolerance is high. The distributed web crawler can craw data from multiple webpage at the same time. In the situation, the probability of program error is high. So, the log monitoring should do well.

In practical system design, not all of system has Web crawler module. Data acquisition module was adopt in [26, 31–34] and there are 3 sentiment analysis systems, 2 text classification systems.

### *4.1 Demands Difference*

The function of classification system is determined by actual requirement. In general, the sentiment analysis has to get data from social media, but, the text classification system has more extensive access to text. Thus, in [34], it is a classification system of call center and all of data to be classified has been stored in the internal database of the unit already. So, Web crawler is useless in the system and the ETL (extraction-transformation-loading) was used to extract text data from internal databases of the unit, so, there is no extranet data in the system and no need to Web crawler. But it is obviously that Web crawler is the mainstream technology of system to get data from outside.

### *4.2 Degree of Perfect About Software is Different*

Metrics to measure web crawlers contain control of download speed, Update in time, avoid repeated crawl and automatic parsing. The Web crawler in [26] can make policies of queue allocation and assign uniform URLs to different queues for ensuring the efficiency of multi process and the ELFHash algorithm was used in here. In addition, the .zip, PDF, Word and.exe files will not be download by the crawler. In [31], the system has a Web crawler that just has basic function and it is single process. In [32], the crawl module can be operated by administrator and MD5 was used to store the text for reducing the amount of memory, garbled and easy storage. In [33], the threading library in python was used to implement the Web crawler and the request header was forged for preventing blocked by server, because, multiple crawl

to same website in a short period of time will trigger the anti-pickpocket mechanism of the website.

So, the basic function of a perfect Web crawler should contain faster download speed, human-computer interaction and Robustness. In addition, it is should be noted that Web crawler module is not the necessary part in classification system, actually, what the system needs is just a data acquisition tool and there are other ways to get data, ETL, for example.

## 5 Relevant and Irrelevant Features

Text preprocessing is the first step in Chinese text classification. In general situation, there are some texts that can be considered as noise in the texts documents [35]. Short text also consists of structured and unstructured data, which must be pre-processed before the classification task is performed [36]. Finding a good approach for data cleaning for conducting analysis smoothly is also important. As the data is collected from various sources, so it contains noise and ambiguity [37]. So, in general situation, the quality of text preprocessing algorithms has a huge influence on the quality of output produced by the text analysis process [38].

**URL Link** Twitter and Weibo platforms allow user to share posts and these include the URL links, @username. Most researchers may consider that the URLs do not contain any information regarding the sentiment of the tweet [39].

**Hashtags** All the private usernames identified by @user and the hashtags identified by the # symbol should be removed.

**Traditional Chinese Texts** In text preprocessing, the traditional Chinese word in short text should be transferred of simplified to simple Chinese word. In addition, few users use traditional Chinese to post in social media. Filtering all the non-Chinese texts, which means all of number, English letter and garbled code in Chinese short text should be removed.

**Emoticons** Emoticons in short text are provided by the Weibo platform for improving sentiment expression. For instance, the emoticons shown in Fig. 3 can be used by users on social media to show some kinds of polarity in the texts for sentiment analysis [40]. Therefore, the emoticons should be reserved in data pre-process.

**Fig. 3** Emoticons

**Word Segmentation** One of the main objectives of text preprocess is to improve accuracy of the word segmentation and the segmentation's accuracy can directly affect the subsequent information extraction from the text. Chinese word segmentation is an important step in NLP and it almost become the standard treatment approach in all of related tasks, such as text retrieval, speech recognition, automatic translation [41]. Furthermore, Chinese words are not separated by space and there is a lack of morphological marker in Chinese sentence. Word segmentation algorithms can be divided into three categories: based on string matching, based on meaning and based on statistics [42].

## 6 Text Features Selection

Text feature extraction can be used to extract text information features and can reveal additional meanings of text [43]. The most common benchmarks used include information gain, mutual information and chi square statistics.

### *6.1 Information Gain (IG)*

Information gain is also called mutual information. It is a kind of method that is based on information theory. The information theory is proposed by Shannon and it uses the method of dividing data set to regularize the disordered data.

### *6.2 Mutual Information*

The degree of correlation between feature item $t$ and category $C_i$ was used to make select about text feature. The value of calculation result is greater, the $t$ and $C_i$ are more related. The formula is as follow:

$$MI(t, C_i) = \log_2 \frac{A*M}{(A+C)*(A+B)} \tag{1}$$

In the formula, $N$ refers to total number of texts in corpus; $A$ is the number of category $C_i$ text that contain $t$; $C$ is the number of category $C_i$ text that do not contain $t$.

### 6.3 Cross Entropy (CE)

Cross entropy also was call KL distance and it represent the distance between probability distribution of the text category and the probability distribution of the text category that a certain feature appear [24]. The related formula is as follow:

$$CE(t) = p(t) \sum_{i=1}^{|c|} p(c_i|t) \log \frac{p(c_i|t)}{p(c_i)} \tag{2}$$

In the formula, $p(t)$ is probability of feature $t$ appearing in text. $p(C_i)$ is the probability of $C_i$ text appearing in the corpus. P($C_i|t$) is the probability of belonging to category C when the text contains feature $t$.

## 7 Calculation of Feature Weights

### 7.1 BOW (Bag-Of-Word)

Bag-of-word is used to express the text feature. In information search, the Word order and grammar was ignored in BOW and just treats them as a set that contains several words. In the text, each word is unrelated, and the appearance of each word are not based on other word. That means, no matter where any word appear in the model, it will not be affected by any other factors [44]. The text analysis will convert these text representations into a numerical form that can be processed by any machine learning (ML) algorithms.

**One-Hot Coding** One-hot coding is also known as one valid code. It uses N-bit status of register to make code to N States. Each state has its independent register bits and only one-bit is effective in any time. In here, the register, small storage area, is used to storage calculation data in CPU [45]. A latch in register can storage 1 binary number, so, N triggers can storage N-bit code of binary. That means that only 1 activation point at one time.

**Term Frequency-Inverse Document Frequency (TF-IDF)** Term frequency means the number of times that a word appears in the text and the number is usually normalized.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

In the formal, $n_{n,j}$ is the appearance time of the word $t_i$ in the text $d_j$. The denominator is appearance time of the word in $d_j$. The IDF was used to measure universal importance of the word. The $tfidf_{i,j} = tf_{i,j} \times idf_i$. Therefore, some high

frequency words in a special document and the frequency of low file that contains the word in entire corpus can generate TF-IDF of high weight.

**Word2vec** Word2vec was generated by Google at 2013 and it is based on deep learning. Its main function is that applying neural network algorithm to seek a continuous space vector for word representation [46]. Word2vec contains CBOW (Continuous Bag of Words) and Skip-Gram. The CBOW likelihood function gives the most likely word which appears with a context. The term skip-gram is an abbreviation for k-skip-n-gram from which n-gram can be a particular case for $k = 0$. In other words, skip-grams are a generalization of n-grams [47].

The above mentioned approaches in text feature extraction were widely used in various classification systems. In [43], the word2vec was used to represent the Chinese Weibo text and the Max pooling layer of CNN as classifier. In related experiment, the accuracy obtained was approximately closed to 96.09%. But, the approach of data washing was not mentioned by the paper. In [6], The main function of feature selection model is that use efficiently ML approach where feature selection techniques are adapted incrementally to select more appropriate concrete features based on healthcare tweets and its corresponding ground truth labels. So, the BOW, TF-IDF and LSI were used to make feature extraction. In related experiment about twitter, the accuracy obtained was closed to 97%.

A more standard process of text preprocess was applied in [36] related to Facebook text and the steps in the paper contain sentence segmentation, word stemming, POS, stop words removal and term weighting, then, the TF-IDF was used in the paper also.

In [39], a comparative study was made about some methods of data preprocess about Twitter text and the result show that The experiments results show that the accuracy and F1-measure can be improved when expanding acronyms and negation replacement were used. But the experiment result barely improves when removing URLs, removing numbers or stop words in data pre-process.

A novel improved Fast Correlation-Based Filter (FCBF [48]) solution was proposed and Feature dimension parameters was introduced into original FCBF in order to have better feature selection algorithm. The related experiments expressed that the performance of the method was better than IG, DFS, FCBF [49, 50].

Next the TF-IDF is a classical algorithm and can be used in feature extraction. Now, some novel methods were proposed such as ant colony optimization, Weight improvement of TF-IDF and Fast Correlation-Based Filter (FCBC) and so on.

## 8 Classifier Module

Classification algorithm is an important module in classification system and the performance of classification system can be determined by it in a large extent.

**Logic Regression** In [20], the logic regression was used to make large-scale short text classification for text matching of FAQ system. In the article, the SVM was used

to make experiment for comparing with regression algorithm and the classification accuracy of it attach to 0.821 (the SVM is 0.809). Actually, the logic regression is a simple and practical, which is borrowed from the field of statistics. The probability of function estimation in LR was used to measure the quantitative relationship between dependent variable and independent variable.

**Support Vector Machine (SVM)** In machine learning method, the SVM algorithm is usually used for making text classification because of its simplicity and easy to be implemented. Particle swarm optimization can be used to improve the SVM performances by tuning loss function, penalty parameter and kernel parameter of SVM [51].

**K-Nearest Neighbors (KNN)** The advantage of KNN algorithm is that running is stably and effectively. Its concept is also simple: Given a test document to be classified, find the most similar document with the classified document. Then, make classification about the document base on the category of the document. So, the $k$ value in the KNN always has great influence on the experimental results. In addition, the disadvantage of KNN contain calculation is large; Prediction speed is relatively slow; So, in [50], a kind of improve KNN was propose base on association rule. The classification algorithm uses the association rule to choose the neighbors and the disadvantage of traditional KNN method are avoided.

**Naive Bayesian (NB)** The model of Naive Bayesian classification is widely used to solve the problem of text classification because of its simplicity, efficiency and effectiveness [49]. The idea of the algorithm is to first calculate the prior probability of each category, then, the posterior probability of each feature belonging to a certain category is calculated by using Bayes theorem and the final category is determined by the maximum estimate of posterior probability.

**Artificial Neural Network** With the development of computer technology, the artificial neural network algorithm gradually becomes one of the important options of text classification technology.

**Convolutional Neural Network (CNN)** CNN is a kind of Feedforward neural network with depth structure. In general, a complete neural network contains input layer, Convolution layer, pooling layer, Full connection layer, output layer.

1. The word vectors of corresponding words in sentence were arranged in sequence (from top to bottom). Let there are number of n word in sentence, dimension of vector is $k$. So, the matrix is $n * k$. The type of matrix can be static or non-static. The word vector in static is fixed. But, in the non-static, it is a parameter that can be optimized, and the process of back error propagation is called as Fine tuning process.
2. In convolution layer, each neuron of the output feature map is locally connected with its input. In terms of its processing, the neural network can be used to extract features more efficient.

3. The pooling layer is used to compress features and remove redundant information. The pooling operation is performed by using the Max Pooling method [52]. So, the main function of pooling layer is for feature dimension reduction and that means only relevant features are extracted automatically.
4. In the full connection layer, the Softmax function was adopted to make the layer become classifier, then the output is the results of the classification process.

**Recurrent Neural Network (RNN)** To be able to use historical information, Recurrent Neural Network (RNN) was introduced [53]. In RNN, the input data at every step not only contains current input data but also contains hidden-layer unit data, which has memory module's function for saving the history information and continue to renew with input new data.

**Long Short Term Memory (LSTM)** A main problem in optimizing RNN that uses gradient descent method is that the gradient may disappear rapidly in the process of back propagation along the sequence [53]. Now, the LSTM was most widely used in practical application for solving the problem of gradient disappearance. In back propagation algorithm, the gradient disappear is not a problem, especially in long-distance task, the LSTM is better than RNN [54]. LSTM is the same as RNN in running mode, but, the different between LSTM and RNN is that a more detailed internal unit was implemented by LSTM and that unit can store information longer. In $t$ time, there are three inputs in LSTM: the input value of the current time, the output value of the previous time and the unit status of previous time. Thus, there are 3 switches to be controlled in the long term status $C$ and they are forgetting gate, input gate and output gate. The above mentioned is a standard LSTM, but in practical application, some improved LSTM models are used in classification system. Furthermore, many kinds of classifier model that have been applied in text classification system also applied in sentiment analysis system.

**Attention-Based** In recent years, attention-based mechanism was proposed, and it has been proven to be very effective in natural language process, such as text translation, text classification and sentence analysis. Attention mechanism can use attention vector to create the degree of relevance assessment between unknown factors and other factors [55]. The attention mechanism can highlight the effect of input on the output, and optimize the traditional model by calculating attention probability distribution [56]. So, in some text classification systems, the attention-based model is used to improve system performance.

Not many works are conducted based on Chinese language. Similarly, not many works published in processing Malay language texts (e.g., Natural language processing for sarcasm [57], sentiment analysis for Chinese texts [58], sentiment analysis using ensemble approach [59], stemming [60, 61], spell checker [62] and part of speech tagger [63]). Table 2 tabulates several works conducted on Chinese corpus. There are 19 papers shown in Table 2. The comparison items contain years, data cleaning, feature extraction, classifier, and accuracy in experiment. The selection criterion is that a novel classification was proposed, or a complete classification system was built. Some of the findings include the following [56]:

**Table 2** Literature review for Chinese sentiment analysis

| Works | Year | Data cleaning | Feature represent | Classifier | Accuracy | Dataset |
|---|---|---|---|---|---|---|
| [4] | 2019 | General processing | Word2Vec | CNN | 95.09% | NLPCC2017/Weibo |
| [15] | 2019 | Word segmentation | TF-IDF | Keyword | NA | Seismic data |
| [33] | 2019 | General processing | Improved TF-IDF | MapReduce | NA | Weibo text |
| [32] | 2019 | General processing | Word2Vec | BiLSTM | NA | NLPCC2017 |
| [28] | 2018 | General processing | TF-IDF | SVM | 95% | 7 Industries theme |
| [50] | 2018 | General processing | TF-IDF*IG | Improved Naïve Bayes | NA | Multiple corpus |
| [64] | 2018 | General processing | TF-IDF | Naïve Bayes | NA | Weibo text |
| [19] | 2017 | General processing | Word2Vec | LSTM | 78–91% | ChnSentiCorp |
| [52] | 2016 | No mentioned | No mentioned | Dirichlet NB | 96% | SouGou text |
| [12] | 2015 | Word segmentation | High-frequency | Naïve Bayes | 100–72% | 10 topics dataset |
| [51] | 2014 | General processing | Ant colony algorithm | KNN | 86% | Recuters-21578 20-Newsgroups |
| [33] | 2013 | Word segmentation | Word2Vec | Naïve Bayes | 65–71% | Weibo text |
| [16] | 2010 | General processing | Multi-level | Distance weight KNN | 87–74% | Subject data of Digital Library |

1. By customizing cyber words during segmentation, more information of Weibo texts can be maintained.
2. By expanding the vocabulary with wiki data, the correlation between word can also be increased.
3. By using k-max pooling method based on length of sentence, more features can also be captured.

## 9 Conclusion and Future Researches

The difficulty of data collection in WeChat account was raised and it has been shown that although WeChat is the largest interactive software in China, but, it is difficult to

get data from the platform for its characteristics of semi closure (limited to acquaintances) [18]. So, there is a problem that the collected data cannot meet the demand of public opinion. Just as mentioned in the paper, the method of data collection is the important research direction of classification system that based on WeChat [18]. In [65], the authors proposed that the timeliness of Web crawler in related system should be improved for analyzing breaking news.

The performance of LDA and TextRank are affected by the number of subjects in Corpus, Number of extracted keywords [19]. Thus, the research direction in the future includes reducing dependence on the influencing factors. Features of social media may include false information or fake news. These kind of false comment, news, information should be removed for ensuring the authenticity of data [26, 35]. An approach of building a joint model was proposed in which the Sentence representation model and Hierarchical classification model should be combined into one model [20]. Thus, the parameters of sentence representation model and parameters of hierarchical classifier can be improved at the same time in each of training phases [1].

Naïve Bayes can be improved by improving the prior probability and model iteration [49]. For instance, designing a model using an improved Naïve Bayes algorithm has been applied in designing flood analysis [4]. The experiment in the paper shows that comprehensive index calculations lead to more reliable values of the prior probability that help improve the estimation of posterior probability and the uncertainty can be reduce in frequency calculation. But its performance in text classification needs to be improved further. A text classification model named NA-CNN-LSTM or NA-CNN-COIF-LSTM has been proposed and the experimental result shows that the proposed model has better performance compared to the standard CNN or LSTM [66]. In the further work of the paper, it was proposed that the combination of CNN and other variants of LSTM should be investigated. The *k*-means clustering algorithm was used in a sentiment analysis system, but, when the data scale is large, the convergence become slow and the response speed of the system is prolonging [64]. So, further works can be explored in improving the *k*-means algorithm for analyzing texts documents. Based on the literature review, the data acquisition, feature extraction and classifier algorithms are the three most important parts that requires improvement.

## References

1. Wadawadagi R, Pagi V (2020) Sentiment analysis with deep neural networks: comparative study and performance assessment. Artif Intell Rev. https://doi.org/10.1007/s10462-020-09845-2
2. C. Team (2019) Weibo monthly active users grew to 462 million in Dec 2018, 93% on mobile [Online]. Available: https://www.chinainternetwatch.com/28566/weibo-fiscal-2018/
3. Yoo S, Song J, Jeong O (2018) Social media contents based sentiment analysis and prediction system 105:102–111
4. Yang X, Xu S, Wu H, Bie R (2019) Sentiment analysis of Weibo comment texts based on extended vocabulary and convolutional neural network. Procedia Comput Sci 147:361–368

5. Sebastiani F, delle Ricérche CN, Moruzi VG (2003) Research in automated text classification: trends and perspectives, no. Ml
6. Kumar S, Kumar S, Suri JS (2019) Computer methods and programs in biomedicine effect of incremental feature enrichment on healthcare text classification system: a machine learning paradigm. Comput Methods Programs Biomed 172:35–51
7. Beier M, Wagner K (2016) Social media adoption: barriers to the strategic use of social media in SMES. Research papers, 100. https://aisel.aisnet.org/ecis2016_rp/100
8. Vashishtha S, Susan S (2019) Fuzzy rule based unsupervised sentiment analysis from social media posts. Expert Syst Appl 138
9. Stieglitz S, Mirbabaie M, Ross B, Neuberger C (2018) Social media analytics—challenges in topic discovery, data collection, and data preparation. Int J Inf Manage 39(October 2017):156–168
10. Wang X, Zhang C, Wu M (2015) Sentiment classification analysis of Chinese microblog network. In: Mangioni G, Simini F, Uzzo S, Wang D (eds) Complex networks VI. Studies in computational intelligence, vol 597. Springer, Cham. https://doi.org/10.1007/978-3-319-16112-9_12
11. Wang J, Lou S, Ming Q (2018) Evaluation of large data analysis function of Chinese national knowledge infrastructure based on mountain tourism research in China. In: 2018 5th International conference on information science and control engineering (ICISCE), Zhengzhou, pp 216–220. https://doi.org/10.1109/ICISCE.2018.00053
12. Gong Z, Yu T (2010) Chinese web text classification system model based on Naive Bayes. In: 2010 International conference on e-product e-service and e-entertainment, Henan, pp 1–4. https://doi.org/10.1109/ICEEE.2010.5660869
13. Ormandjieva O, Hussain I, Kosseim L (2007) Toward a text classification system for the quality assessment of software requirements written in natural language, pp 39–45. https://doi.org/10.1145/1295074.1295082
14. Simanjuntak D, Purnomo Ipung H, Lim C, Nugroho A (2011) Text classification techniques used to facilitate cyber terrorism investigation, pp 198–200. https://doi.org/10.1109/ACT.2010.40
15. Simon T, Goldberg A, Adini B (2015) Socializing in emergencies—a review of the use of social media in emergency situations. Int J Inf Manage 35(5):609–619. ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2015.07.001
16. Yi K (2007) Automated text classification using library classification schemes: trends, issues, and challenges. Int Cataloguing Bibliographic Control J 36:78–82
17. Yang L, Li Y, Wang J, Sherratt RS (2020) Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access 8:23522–23530. https://doi.org/10.1109/ACCESS.2020.2969854
18. Zhang S, Shang K, Cong S, Zhang B, Liu Z (2012) WIPOMTS: an internet public opinion monitoring system. In: Liu C, Wang L, Yang A (eds) Information computing and applications. ICICA 2012. Communications in computer and information science, vol 307. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34038-3_1
19. Wong S, Dastane O, Mohd Satar N, Ma'arif M (2019) What WeChat can learn from WhatsApp? Customer value proposition development for mobile social networking (MSN) apps: a case study approach. J Theor Appl Inf Technol 97:1091–1117
20. Kim K-M, Kim Y, Lee J, Lee J-M, Lee SK (2019) From small-scale to large-scale text classification. In: The World Wide Web conference (WWW'19). Association for Computing Machinery, New York, NY, USA, pp 853–862. https://doi.org/10.1145/3308558.3313563
21. Kim K-H (2011) Design and implementation of opinion mining system based on association model. J Korean Inst Inf Commun Eng 15:133–140. https://doi.org/10.6109/jkiice.2011.15.1.133
22. Faed A (2013) An intelligent customer complaint management system with application to the transport and logistics industry. https://doi.org/10.1007/978-3-319-00324-5
23. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113. ISSN 2090-4479, https://doi.org/10.1016/j.asej.2014.04.011

24. Alfred R, Teoh RW (2019) Improving topical social media sentiment analysis by correcting unknown words automatically. In: Yap B, Mohamed A, Berry M (eds) Soft computing in data science. SCDS 2018. Communications in computer and information science, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-3441-2_23
25. Hung P, Lai, Rayner, Alfred (2019) An optimized multi-layer ensemble framework for sentiment analysis. In: 2019 1st International conference on artificial intelligence and data sciences (AiDAS), Ipoh, Perak, Malaysia, pp 158–163. https://doi.org/10.1109/AiDAS47888.2019.8970949
26. Brynielsson J, Johansson F, Jonsson C et al (2014) Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. Secur Inf 3:7. https://doi.org/10.1186/s13388-014-0007-3
27. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012) A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In: Proceedings of the 50th annual meeting of the association for computational linguistics, July, pp 115–120
28. Miao F, Zhang P, Jin L, Wu H (2018) Chinese news text classification based on machine learning algorithm. In: 2018 10th International conference on intelligent human-machine systems and cybernetics, vol 02, pp 48–51
29. Kumar N, Singh M (2016) Framework for distributed semantic web crawler. In: Proceedings—2015 International conference on computational intelligence and communication networks, CICN 2015, pp 1403–1407
30. Bal S (2012) The issues and challenges with the web crawlers. Int J Inf Technol Syst 1:1–10
31. Drus Z, Khalid H (2019) Sentiment analysis in social media and its application: systematic literature review. Procedia Comput Sci 161:707–714. ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.11.174
32. Liu J, Chang W-C, Wu Y, Yang Y (2017) Deep learning for extreme multi-label text classification. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval (SIGIR'17). Association for Computing Machinery, New York, NY, USA, pp 115–124. https://doi.org/10.1145/3077136.3080834
33. Li X, Gao L (2013) The design and implementation of an internet public opinion monitoring and analyzing system, pp 176–180. https://doi.org/10.1109/ICSS.2013.59
34. Tang M, Pellom B, Hacioglu K (2003) Call-type classification and unsupervised training for the call center domain, pp 204–208. https://doi.org/10.1109/ASRU.2003.1318429
35. Liang H, Sun X, Sun Y, Gao Y (2017) Text feature extraction based on deep learning: a review. EURASIP J Wirel Commun Networking 2017(1):211. https://doi.org/10.1186/s13638-017-0993-1
36. Sriyanong W, Moungmingsuk N, Khamphakdee N (2018) A text preprocessing framework for text mining on big data infrastructure. In: 2018 2nd International conference on imaging, signal processing and communication, ICISPC 2018, pp 169–173
37. Kumar V, Khosla C (2018) Data cleaning—a thorough analysis and survey on unstructured data. In: Proceedings of the 8th International Conference on Confluence. 2018 Cloud Computing and Data Science Engineering, pp 305–309
38. Chandrasekar P, Qian K (2016) The impact of data preprocessing on the performance of a Naïve Bayes classifier. In: Proceedings of the international computer software and applications conference, vol 2, pp 618–619
39. Jianqiang Z, Xiaolin G (2017) Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access 5:2870–2879
40. Wang H, Castanon JA (2015) Sentiment expression via emoticons on social media. In: Proceedings—2015 IEEE international conference on big data. IEEE big data 2015, pp 2404–2408
41. Chen W, Chen B, Xiang T, Zhang Z (2010) A pragmatic approach to increase accuracy of Chinese word-segmentation. In: Proceedings—2010 International forum on information technology and applications, IFITA 2010, vol 1, pp 389–391
42. Altınel B, Ganiz MC (2018) Semantic text classification: a survey of past and recent advances. Inf Process Manage 54(6):1129–1153. ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2018.08.001

43. Abuzaraida MA, Zeki AM, Zeki AM (2013) Feature extraction techniques of online handwriting arabic text recognition. In: 2013 5th International conference on information and communication technology for the Muslim World (ICT4M), Rabat, 2013, pp 1–7. https://doi.org/10.1109/ICT4M.2013.6518884
44. Zhao P, Cai Q-S, Wang Q-Y, Geng H-T (2007) Automatic keyword extraction of Chinese document algorithm based on complex network features 20:827–831
45. Yu M (2013) An 8 bit 12 MS/s asynchronous successive approximation register ADC with an on-chip reference. Chin Inst Electron J Semicond 34(2). https://doi.org/10.1088/1674-4926/34/2/025010
46. Wan S, Li B, Zhang A, Wang K, Li X (2018) Vertical and sequential sentiment analysis of micro-blog topic. In: Gan G, Li B, Li X, Wang S (eds) Advanced data mining and applications. ADMA 2018. Lecture notes in computer science, vol 11323. Springer, Cham. https://doi.org/10.1007/978-3-030-05090-0_30
47. Song M, Yoo CD (2016) Multimodal representation: Kneser-ney smoothing/skip-gram based neural language model. In: 2016 IEEE International conference on image processing (ICIP), Phoenix, AZ, pp 2281–2285. https://doi.org/10.1109/ICIP.2016.7532765
48. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings, twentieth international conference on machine learning, vol 2, pp 856–863
49. Li-Guo D, Peng D, Ai-Ping L (2014). A new Naive Bayes text classification algorithm. TELKOMNIKA Indonesian J Electr Eng 12. https://doi.org/10.11591/telkomnika.v12i2.4180
50. Wang L, Zhao X (2012) Improved KNN classification algorithms research in text categorization. https://doi.org/10.1109/CECNet.2012.6201850
51. Wang L, Mu X, Liu H (2020) Using SVM method optimized by improved particle swarm optimization to analyze emotion of Chinese text. Comput Sci 47(1):231–236
52. Ganda R, Mahmood A (2017) Efficient deep learning model for text classification based on recurrent and convolutional layers. https://doi.org/10.1109/ICMLA.2017.00009
53. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Networks
54. Gers FA, Schraudolph NN, Schmidhuber J (2002) Learning precise timing with LSTM recurrent networks. J Mach Learn Res
55. Tang X, Chen Y (2019) A multi-scale convolutional attention based GRU network for text classification, pp 3009–3013
56. Bai X (2018) Text classification based on LSTM and attention. In: 2018 Thirteenth international conference on digital information management (ICDIM), Berlin, Germany, pp 29–32. https://doi.org/10.1109/ICDIM.2018.8847061
57. Leong LC, Basri S, Alfred R (2012) Enhancing Malay stemming algorithm with background knowledge. In: Anthony P, Ishizuka M, Lukose D (eds) PRICAI 2012: trends in artificial intelligence. PRICAI 2012. Lecture notes in computer science, vol 7458. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32695-0_68
58. Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F (2017) Natural language processing based features for sarcasm detection: an investigation using bilingual social media texts. In: 2017 8th International conference on information technology (ICIT), Amman, pp 703–709. https://doi.org/10.1109/ICITECH.2017.8079931
59. Wang D, Alfred R (2020) A review on sentiment analysis model for Chinese Weibo text. In: 2020 3rd International conference on advanced electronic materials, computers and software engineering (AEMCSE), Shenzhen, China, pp 456–463. https://doi.org/10.1109/AEMCSE50948.2020.00105
60. Basri S, Alfred R, On CK (2012) Automatic spell checker for Malay blog. In: 2012 IEEE International conference on control system, computing and engineering, Penang, pp 506–510. https://doi.org/10.1109/ICCSCE.2012.6487198
61. Alfred R, Leong LC, On CK, Anthony P (2014) A literature review and discussion of malay rule—based affix elimination algorithms. In: Uden L, Wang L, Corchado Rodríguez J, Yang HC, Ting IH (eds) The 8th international conference on knowledge management in organizations.

Springer proceedings in complexity. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7287-8_23
62. Alfred R, Mujat A, Obit JH (2013) A ruled-based part of speech (RPOS) Tagger for Malay text articles. In: Selamat A, Nguyen NT, Haron H (eds) Intelligent information and database systems. ACIIDS 2013. Lecture notes in computer science, vol 7803. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36543-0_6
63. Luan Y, Lin S (2019) Research on text classification based on CNN and LSTM. In: 2019 IEEE International Conference on Artificial Intelligence and Computer Applications, pp 352–355
64. Li Y, Zhou X, Sun Y, Zhang H (2016) Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. China Commun 13:91–105. https://doi.org/10.1109/CC.2016.7781721
65. Miao F, Zhang P, Jin L, Wu H (2018) Chinese news text classification based on machine learning algorithm, pp 48–51. https://doi.org/10.1109/IHMSC.2018.10117
66. Wang Y, Gao P, Guo A, Chang J, Zhao M (2019) Application of Bayesian model with improved prior probability in design flood analysis. Shuili Fadian Xuebao/J Hydroelectr Eng 38(7):67–76

# Newton-SOR with Quadrature Scheme for Solving Nonlinear Fredholm Integral Equations


**L. H. Ali, J. Sulaiman, A. Saudi, and M. M. Xu**

**Abstract** This paper presents a numerical method of Newton Successive Over-Relaxation (NSOR) iteration with quadrature scheme to approximate the solution of nonlinear Fredholm integral equations. Here, the quadrature scheme is used to derive the approximation equations of nonlinear Fredholm integral equations in order to develop a system of nonlinear equations. NSOR consists of two parts. In the first part, Newton's method is used to linearize the developed system of nonlinear equations. Then, in the second part, SOR iteration is used to solve the corresponding system of linear equations to get the approximate solution. In order to validate the performance of the proposed method, Newton-Jacobi (NJacobi) and Newton-Gauss–Seidel (NGS) are used as the reference methods to perform the comparative analysis. Also, some numerical examples are presented to illustrate the validity of the NSOR.

**Keywords** Nonlinear fredholm integral equations · Quadrature scheme · Newton-SOR · Trapezium rule · Newton's method

L. H. Ali (✉) · J. Sulaiman · M. M. Xu
Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, 88400 Sabah, Malaysia
e-mail: labiyana15@gmail.com

J. Sulaiman
e-mail: jumat@ums.edu.my

M. M. Xu
e-mail: xmmzg@sina.com

A. Saudi
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, 88400 Sabah, Malaysia
e-mail: azali@ums.edu.my

M. M. Xu
School of Mathematics and Information Technology, Xingtai University, 88, Quanbei East Street, Xingtai City, China



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_27

## 1 Introduction

Nonlinear integral equations appear in many scientific fields such as fluid mechanics, biological models, solid state physics, and kinetics chemistry [1]. The nonlinear integral equations can be solved either analytically or numerically. However, analytical approach requires tedious calculation and not it is easy to be applied as many definite integrals need to be computed in solving nonlinear integral equations [2]. Therefore, abundant scientific research has been conducted to develop numerical methods to solve the nonlinear Fredholm integral equations. The latest research on solving nonlinear Fredholm integral equations involves several numerical methods such as the Adomian decomposition method [3], multi-projection method [4], Romberg quadrature rule [5], successive approximation method [6], Nystrom method [7], Nystrom-quasi-linearization method [8], and cosine-trigonometric approximation method [9].

In this paper, we attempted to apply the NSOR iteration with quadrature scheme to extract the approximate solutions of the following nonlinear Fredholm integral equations of the second kind in the following form:

$$u(x) = g(x) + \int_a^b k(x, y, u(y))dy, \; y \in [a, b], \tag{1}$$

where $k$ is continuous on $[a, b]$, $g(x)$ is known function, and $u(x)$ is the unknown function [10]. The main idea of this study is to apply the quadrature scheme to generate approximation equations of Eq. (1). Then, we used it to develop a system of nonlinear equations. After that, we applied Newton's method to linearize the developed system of nonlinear equations to a system of linear equations. Here, a family of weighted iterative methods such as SOR [11–13] and Accelerated Overrelaxation (AOR) method [14, 15] can be considered a linear solver in solving any large linear system. However, this study deals with the implementation of SOR iteration to solve the corresponding system of linear equations to get the approximate solution of nonlinear Fredholm integral equations. The concept of SOR is to add the weighted parameter in iteration to produce faster convergence and more efficient point iterative method compared to Jacobi and Gauss–Seidel [11].

NSOR iteration is the combination of Newton's method and SOR iterative method. NSOR iteration is the extended method from NGS that produces a faster convergence method to solve any nonlinear systems [16]. In general, the system of nonlinear equations is difficult to be solved especially when involving a large-scale system. The advantage of NSOR is that the large-scale system of linear equations can be generated using linearization then it can be solved iteratively using SOR iterative method which only requires minimum amount of mathematical operations. In the previous studies, the discussion such as in [3–9] involves a small number of node points. Thus, in this study, we will present the implementation of the proposed method

on high order matrices by considering the number of subinterval, $n$ to be 256, 512, 1024, 2048 and 4096.

This work consists of several sections. In the next section, we will discuss the methodology used in this study to generate a system of nonlinear equations. Then, we will apply Newton's method to represent the system of nonlinear equations in a linear form. After that, we will discuss the application of the proposed method to solve the generated system of linear equations in the following sub-section. In Sect. 3, we will present the numerical examples that will be used to illustrate the efficiency of the proposed method. Then, we will discuss the numerical results obtained from the numerical experiments. Finally, we will conclude our findings in Sect. 4.

## 2 Methodology

In this section, we will discuss the formulation of NFIE-2 using the discretization scheme to generate the corresponding system of nonlinear equations. Note that in this study, we used the first-order quadrature scheme which is the Trapezium rule. Besides that, we will also discuss the implementation of Newton's method to represent the generated system of nonlinear equations in a linear form. After the linear system is formed, we will continue the discussion by discussing the formulation of NSOR iteration to solve the system of linear equations.

### *2.1 Discretization of Nonlinear Fredholm Integral Equations Using Quadrature Scheme*

Consider the nonlinear Fredholm integral equations in Eq. (1). Suppose.

$$u(x) - \int_a^b k(x, y, u(y))dy, = g(x) \tag{2}$$

and assume that $k(x, y, u(y))$ is a nonlinear function of $u(x)$. This means that Eq. (2) contains the nonlinear function presented by $k(x, y, u(y))$. Let the interval $[a, b]$ be the uniformly partition interval as shown in Fig. 1, so

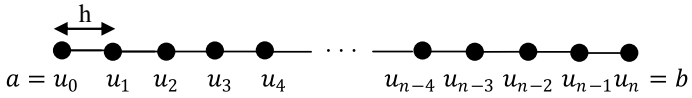



**Fig. 1** Distribution of node points on interval $[a, b]$

$$\int_a^b k(x, y, u(y))dy = \int_{u_i}^{u_n} k(x, y, u(y))dy, \tag{3}$$

where $i = 0, 1, 2, \ldots, n$.

By using quadrature scheme of Trapezium rule, we can consider all the node points in Fig. 1 and form the corresponding nonlinear approximation equations of [12]

$$u_i - \frac{1}{2}hk(x, y_0, u_0) - hk(x, y_1, u_1) - hk(x, y_2, u_2) - \cdots - \frac{1}{2}hk(x, y_n, u_n) = g_i. \tag{4}$$

One of the advantages of quadrature scheme on integral equations is that when the number of subinterval, $n$ increases, the total number of node points will increase as well, thus resulting the neighboring distance of each node point on interval $[a, b]$ to be increased. Therefore, the larger $n$ is expected to give more accurate results for the problem of integral equations.

The nonlinear function of Eq. (4) can be represented in the expression given [17]:

$$\begin{aligned} F_i(u_0, u_1, u_2, \ldots, u_n) = u_i - \frac{1}{2}hk(x, y_0, u_0) - hk(x, y_1, u_1) \\ - hk(x, y_2, u_2) - \cdots - \frac{1}{2}hk(x, y_n, u_n) - g_i. \end{aligned} \tag{5}$$

Then, we can form the corresponding nonlinear system based on Eq. (5) as follows:

$$F_i(u_0, u_1, u_2, \ldots, u_n) = 0, \tag{6}$$

Now, we use the Newton's method to linearize the developed system of nonlinear equations in Eq. (6) to generate a system linear equation in the given expression [16, 17]:

$$J\left(\underline{u}^{(k)}\right)\nabla\underline{u}^{(k)} = -F\left(\underline{u}^{(k)}\right), \tag{7}$$

where $J\left(\underline{u}^{(k)}\right)$ is the Jacobian matrix, $\nabla\underline{u}$ is the vector, and the solution vector is determined by the following formulation:

$$u_i^{(k+1)} = u_i^{(k)} + \nabla u_i, i = 0, 1, 2, \ldots, n. \tag{8}$$

In terms of simplicity, the system of linear Eq. (7) is now rewritten in the following form:

$$A\nabla\underline{u} = \underline{b}, \tag{9}$$

where

$$A = \begin{bmatrix} A_{0,0} & A_{0,1} & A_{0,2} \cdots A_{0,n} \\ A_{1,0} & A_{1,1} & A_{1,2} \cdots A_{1,n} \\ A_{2,0} & A_{2,1} & A_{2,2} \cdots A_{2,n} \\ \vdots & \vdots & \vdots \ddots \vdots \\ A_{n,0} & A_{n,1} & A_{n,2} \cdots A_{n,n} \end{bmatrix}_{(n+1)\times(n+1)},$$

$$\nabla \underline{u} = [\nabla u_0, \nabla u_1, \nabla u_2, \ldots, \nabla u_n]^T,$$

$$\underline{b} = [b_0, b_1, b_2, \ldots, b_n]^T.$$

## 2.2 Newton-SOR Iteration

The SOR iteration method is a well-known iterative method with its weighted parameter which provides more efficient results by accelerating the convergence rate of iteration. To implement this SOR iteration on the system of linear Eq. (1), the coefficient matrix $A$ in Eq. (9) needs to be decomposed into $A = D - L - U$ where the formulation of SOR in solving the NFIE-2 is given as follows [11–13]

$$\nabla u^{(k+1)} = T_{\text{SOR}} \nabla u^{(k)} + (D - \omega L)^{-1} b, \tag{10}$$

$$T_{\text{SOR}} = (D - \omega L)^{-1}((1 - \omega)D + \omega U), \omega \in (0, 2), \tag{11}$$

with $D, L$ and $U$ indicating the diagonal matrix, lower triangular matrix and upper triangular matrix of coefficient matrix $A$ respectively. The algorithm of NSOR iteration is shown as follows.

**Algorithm 1** Implementation of NSOR Iteration.

i. Let $\nabla \underline{u}^{(k)} = 0$ and $k = 0$ be the initial value and $\varepsilon = 10^{-10}$.
ii. Assign $q = 0$ and compute
    a. $A = J(\underline{u}^{(k)})$
    b. $\underline{b} = -F(\underline{u}^{(k)})$
iii. Find the current value, $\nabla \underline{u}^{(k+1)}$
    a. For $i = 0, 1, 2, \ldots, n$, calculate

$$\nabla u_i^{(k+1)} \leftarrow \begin{cases} \left(\nabla u_i^{(k)} + \frac{\omega}{A_i K_{i,i}}\left(b_i - \sum_{j=1}^{n} A_i K_{i,i} \nabla u_j^{(k)}\right)\right), & i = 0 \\ \left(\nabla u_i^{(k)} + \frac{\omega}{A_i K_{i,i}}\left(b_i - \sum_{j=0}^{n-1} A_i K_{i,i} \nabla u_j^{(k+1)}\right)\right), & i = n \\ \left(\nabla u_i^{(k)} + \frac{\omega}{A_i K_{i,i}}\left(b_i - \sum_{j=0}^{n-1} A_i K_{i,i} \nabla u_j^{(k+1)} - \sum_{j=1}^{n} A_i K_{i,i} \nabla u_j^{(k)}\right)\right), & otherwise \end{cases}.$$

b. Repeat step iii(a) until the convergence criterion $\left|\nabla u_i^{(k+1)} - \nabla u_i^{(k)}\right| \leq \varepsilon$ is satisfied, otherwise go to step iv.

iv. Repeat step ii until the convergence criterion $\left|\nabla u_i^{(k+1)} - \nabla u_i^{(k)}\right| \leq \varepsilon$ is satisfied, otherwise go to step v.
v. Display the output.
vi. Stop.

## 3 Numerical Examples and Discussions

In this section, we will present four numerical examples from the previous studies in [10, 18–20] to illustrate the efficiency of the proposed method. All the numerical results obtained are presented in tables. Then, we will discuss the numerical results obtained from the numerical experiments.

### *3.1 Numerical Examples*

In order to illustrate the efficiency of the NSOR, we will use NJacobi and NGS as the reference methods. Also, to show the validity of the methodology discussed in the previous section, all of these methods will be tested on the following numerical examples of nonlinear Fredholm integral equations of the second kind.

**Example 1** In this example, we consider the following equation [10]:

$$u(x) = x + \frac{\cos\left(e^{(1)} + x\right) - \cos(1 + x)}{20} + \int_0^1 \frac{\sin\left(e^{(y)} + x\right)}{20} e^{(u(y))} dy. \tag{12}$$

The exact solution of this integral equation is $u(x) = x$.

**Example 2** In this example, we consider the following equation [18]:

$$u(x) = 1 - \frac{5}{12}x + \int_0^1 xy[u(y)]^2 dy. \tag{13}$$

The exact solution of this integral equation is $u(x) = 1 + \frac{1}{3}x$.

**Example 3** In this example, we consider the following equation [19]:

$$u(x) = -\frac{x}{9} - \frac{x^2}{8} + x^3 + \int_0^1 \left(x^2 y + xy^2\right) u^2(y) dy. \tag{14}$$

The exact solution of this integral equation is $u(x) = x^3$.

**Example 4** In this example, we consider the following equation [20]:

$$u(x) = \left(\frac{1}{2} - ln(2)\right)x^2 + \sqrt{x} + \int_0^1 \frac{x^2 y^2}{1 + u^2(y)} dy. \tag{15}$$

The exact solution of this integral equation is $u(x) = \sqrt{x}$.

## *3.2 Discussions*

Based on Algorithm 1, we have conducted the experimental test using C language on Borland C++ version 5.02 on several numbers of subintervals, $n$, which are 256, 512, 1024, 2048, and 4096. Then, a comparative analysis between NSOR with the reference methods was done by using three parameters; number of iterations (I), computational time (Time) measured in seconds, and maximum absolute error (Error). The numerical results from the implementation of NJacobi, NGS, and NSOR on Example 1 to 4 are tabulated in Tables 1, 2, 3 and 4, respectively.

Based on these tables, it is clearly shows that the implementation of NSOR iteration only requires the smallest number of iteration and fastest computational time compared to NJacobi and NGS iteration. For the comprehensive analysis, we have presented the data in Table 5. Based on Table 5, NSOR method recorded bigger of reduction percentages compared to NGS iteration in terms of number of iteration and computational time. It also shows that the NSOR iterative method can reduce the number of iterations approximately 27.27%, 63.53–63.64%, 48.57% and 46.58% for Example 1, 2, 3 and 4 respectively. In terms of computational time, NSOR method can reduce the computation time approximately 22.49–23.14% (Example 1), 60.84–62.22% (Example 2), 41.82–46.42% (Example 3) and 37.50–41.35% (Example 4) compared to NJacobi. For the maximum absolute error, the data has been transform into a graphical form for further discussion as follows.

**Table 1** Numerical results of NJacobi, NGS, and NSOR for Example 1

| $n$ | Method | $I$ | Time | Error |
|---|---|---|---|---|
| 256 | NJacobi | 22 | 0.43 | 5.81294E-07 |
| | NGS | 18 | 0.36 | 5.81294E-07 |
| | NSOR | 16 | 0.33 | 5.81294E-07 |
| 512 | NJacobi | 22 | 1.69 | 1.45324E-07 |
| | NGS | 18 | 1.42 | 1.45324E-07 |
| | NSOR | 16 | 1.31 | 1.45324E-07 |
| 1024 | NJacobi | 22 | 6.73 | 3.63309E-08 |
| | NGS | 18 | 5.66 | 3.63309E-08 |
| | NSOR | 16 | 5.21 | 3.63309E-08 |
| 2048 | NJacobi | 22 | 26.94 | 9.08274E-09 |
| | NGS | 18 | 22.72 | 9.08272E-09 |
| | NSOR | 16 | 20.76 | 9.08272E-09 |
| 4096 | NJacobi | 22 | 108.01 | 2.27070E-09 |
| | NGS | 18 | 90.89 | 2.27068E-09 |
| | NSOR | 16 | 83.02 | 2.27068E-09 |

**Table 2** Numerical results of NJacobi, NGS, and NSOR for Example 2

| $n$ | Method | $I$ | Time | Error |
|---|---|---|---|---|
| 256 | NJacobi | 329 | 0.45 | 1.27161E-05 |
| | NGS | 182 | 0.26 | 1.27164E-05 |
| | NSOR | 120 | 0.17 | 1.27165E-05 |
| 512 | NJacobi | 329 | 1.66 | 3.17858E-06 |
| | NGS | 183 | 0.94 | 3.17884E-06 |
| | NSOR | 120 | 0.65 | 3.17895E-06 |
| 1024 | NJacobi | 330 | 6.65 | 7.94340E-07 |
| | NGS | 183 | 3.76 | 7.94608E-07 |
| | NSOR | 120 | 2.52 | 7.94720E-07 |
| 2048 | NJacobi | 330 | 26.56 | 1.98289E-07 |
| | NGS | 183 | 14.97 | 1.98558E-07 |
| | NSOR | 120 | 10.04 | 1.98670E-07 |
| 4096 | NJacobi | 330 | 106.12 | 4.92759E-08 |
| | NGS | 184 | 60.23 | 4.95456E-08 |
| | NSOR | 120 | 40.20 | 4.96584E-08 |

**Table 3** Numerical results of NJacobi, NGS, and NSOR for Example 3

| $n$ | Method | $I$ | Time | Error |
|---|---|---|---|---|
| 256 | NJacobi | 105 | 0.16 | 4.50768E-05 |
| | NGS | 64 | 0.11 | 4.50768E-05 |
| | NSOR | 54 | 0.09 | 4.50768E-05 |
| 512 | NJacobi | 105 | 0.55 | 1.12683E-05 |
| | NGS | 64 | 0.35 | 1.12683E-05 |
| | NSOR | 54 | 0.32 | 1.12683E-05 |
| 1024 | NJacobi | 105 | 2.19 | 2.81697E-06 |
| | NGS | 64 | 1.41 | 2.81702E-06 |
| | NSOR | 54 | 1.18 | 2.81703E-06 |
| 2048 | NJacobi | 105 | 8.78 | 7.04191E-07 |
| | NGS | 64 | 5.54 | 7.04247E-07 |
| | NSOR | 54 | 4.72 | 7.04254E-07 |
| 4096 | NJacobi | 105 | 35.05 | 1.76000E-07 |
| | NGS | 64 | 22.00 | 1.76056E-07 |
| | NSOR | 54 | 18.78 | 1.76063E-07 |

**Table 4** Numerical results of NJacobi, NGS, and NSOR for Example 4

| $n$ | Method | $I$ | Time | Error |
|---|---|---|---|---|
| 256 | NJacobi | 73 | 0.32 | 8.60252E-07 |
| | NGS | 43 | 0.21 | 8.60251E-07 |
| | NSOR | 39 | 0.20 | 8.60251E-07 |
| 512 | NJacobi | 73 | 1.33 | 2.15064E-07 |
| | NGS | 43 | 0.79 | 2.15063E-07 |
| | NSOR | 39 | 0.78 | 2.15063E-07 |
| 1024 | NJacobi | 73 | 5.01 | 5.37669E-08 |
| | NGS | 43 | 3.08 | 5.37658E-08 |
| | NSOR | 39 | 2.98 | 5.37658E-08 |
| 2048 | NJacobi | 73 | 19.99 | 1.34425E-08 |
| | NGS | 43 | 12.33 | 1.34414E-08 |
| | NSOR | 39 | 12.27 | 1.34413E-08 |
| 4096 | NJacobi | 73 | 79.33 | 3.36138E-09 |
| | NGS | 43 | 49.51 | 3.36028E-09 |
| | NSOR | 39 | 46.62 | 3.36022E-09 |

**Table 5** Comparison of reduction in percentages for NSOR and NGS compared with NJacobi

| Method | Number of iterations (%) | | | |
|---|---|---|---|---|
| | Example 1 | Example 2 | Example 3 | Example 4 |
| NGS | 18.18 | 44.24–44.68 | 39.05 | 41.10 |
| NSOR | 27.27 | 63.53–63.64 | 48.57 | 46.58 |
| Methods | Computational time (%) | | | |
| | Example 1 | Example 2 | Example 3 | Example 4 |
| NGS | 15.66–16.28 | 42.227–43.64 | 31.25–37.23 | 39.32–41.35 |
| NSOR | 22.49–23.14 | 60.84–62.22 | 41.82–46.42 | 37.50–41.35 |

Figures 2, 3, 4 and 5 shows the behaviour of maximum absolute error for all tested methods towards the matrix sizes for Example 1, 2, 3 and 4 respectively. These figures show the implementation of all methods on numerical examples is in a good agreement. Besides that, the maximum absolute error for all examples shows a very accurate results as the value of $n$ is increased. This is due to the diminution

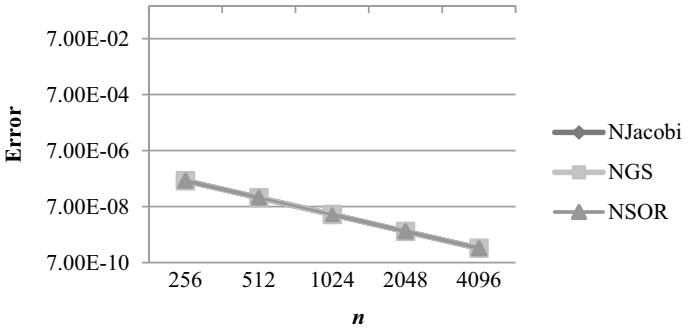



**Fig. 2** Comparison of maximum absolute error towards $n$ for Example 1

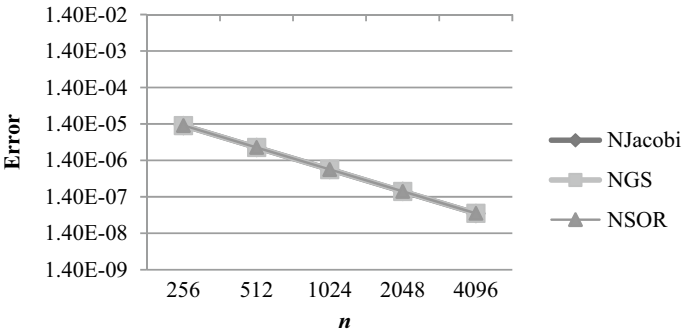



**Fig. 3** Comparison of maximum absolute error towards $n$ for Example 2

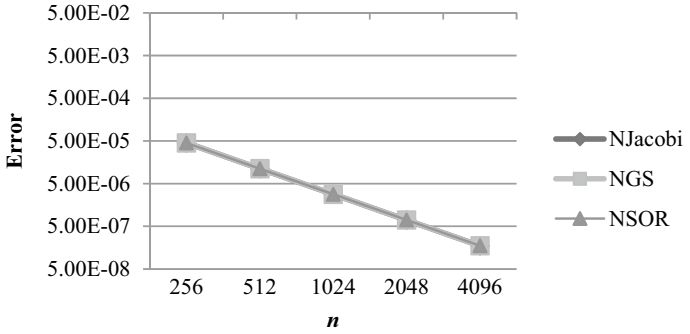



**Fig. 4** Comparison of maximum absolute error towards $n$ for Example 3

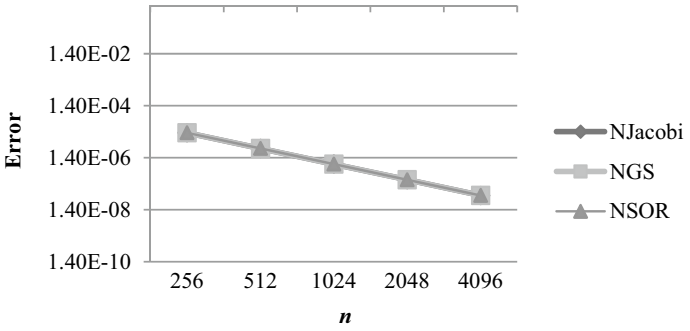



**Fig. 5** Comparison of maximum absolute error towards $n$ for Example 4

of neighbouring distance between each node points on interval $[a, b]$. Thus, the size 4096 shows the most accurate results compared to the other tested number of $n$.

## 4 Conclusions

From the numerical experiment, the NJacobi, NGS and NSOR iterative methods have been applied successfully in finding the approximate solution of nonlinear Fredholm integral equations. Based on the result obtained, we can conclude that NSOR is superior in terms of iteration number and computational time compared to reference methods. In terms of maximum absolute error, the result for all methods shows a good agreement. Also, we can conclude that the maximum absolute error for all examples approaching zero as the order of matrices is increasing due to the diminution of neighboring distance between each node points on interval $[a, b]$. Therefore, the larger grid size gives more accurate results. Besides that, all the iterative methods which have been discussed in this study are categorized in a family of full-sweep

iteration where the implementation of iterative methods is used to calculate all the node points on interval $[a, b]$. Thus, in future research, this research finding could be extended by using the half-sweep iteration method such as in [21–23], where the combination of iterative method and direct method can be taking placed to increase the convergence rate.

**Acknowledgements** We would like to thank the anonymous reviewers of this paper for their suggestions in improving the presentation of this paper. We also would like to show our gratitude to Centre for Postgraduate Studies, Universiti Malaysia Sabah for funding this conference fee. This research is funded by Universiti Malaysia Sabah under the UMS Research Grant Scheme, GUG0485-1/2020.

## References

1. Li H, Huang J (2016) A novel approach to solve nonlinear Fredholm integral equations of the second kind. SpringerPlus 5(154):1–9
2. Borzabadi AH, Fard OS (2009) Numerical scheme for a class of nonlinear Fredholm integral equations of the second kind. J Comput Appl Math 232:449–454
3. Mohedul HMd, Abdul MMd (2017) Solving nonlinear integral equations by using Adomian decomposition method. J Appl Comput Math 6(2):1–4
4. Das P, Nelakanti G (2018) Error analysis of polynomial-based multi-projection methods for a class of nonlinear fredholm integral equations. J Appl Math Comput 56:1–24
5. Katani R (2018) Numerical solution of the Fredholm integral equations with a quadrature method. SeMA J 76:271–276
6. Maturi DA (2019) The successive approximation method for solving nonlinear Fredholm integral equation of the second kind using maple. Adv Pure Math 9:832–843
7. Awawdeh F, Smail L (2020) Convergence analysis of a highly accurate nystrom scheme for fredholm integral equations. Appl Numerical Math 152:231–242
8. Najafi E (2020) Nystrom-quasilinearization method and smoothing transformation for the numerical solution of nonlinear weakly singular Fredholm integral equations. J Comput Appl Math 368:1–13
9. Amiri S, Hajipour M, Baleanu D (2020) On accurate solution of the Fredholm integral equations of the second kind. Appl Numerical Math 150:478–490
10. Allahviranloo T, Ghanbari M (2011) Discrete homotopy analysis method for the nonlinear Fredholm integral equations. Ain Shams Eng J 2:133–140
11. Young DM (1954) Iterative methods for solving partial difference equations of elliptic type. Trans Am Math Soc 76(1):92–111
12. Ali LH, Sulaiman J, Hashim SRM (2018) SOR iterative method with Simpson's 1/3 rule for the numerical solution of fuzzy second kind Fredholm integral equations. In: Journal of Physics: Conference Series, International Conference on Fundamental & Applied Sciences, A Conference of World Engineering, Science & Technology Congress. IOP Publishing Ltd, Kuala Lumpur, Malaysia, pp 1–9
13. Ali LH, Sulaiman J, Hashim SRM (2018) Numerical solution of SOR iterative method for fuzzy Fredholm integral equations of second kind. In: Proceeding of the international conference on mathematics, engineering and industrial applications. AIP conference Proceedings, Kuala Lumpur, Malaysia, pp 1–8
14. Hadjidimos A (2000) Successive Overrelaxation (SOR) and related methods. J Comput Appl Math 123:177–199

15. Sunarto A, Sulaiman J, Saudi A (2014) Implicit finite difference solution for time-factional diffusion equations using AOR method. In: Journal of Physics: Conference Series, Volume 495, 2014 International Conference on Science & Engineering in Mathematics, Chemistry and Physics. IOP Publishing Ltd, Jakarta, Indonesia, pp 1–9
16. Chew JVL, Sulaiman J (2017) Newton-SOR iterative method for solving the two-dimensional porous medium equations. J Fundamental Appl Sci 9(6S):384–394
17. Sulaiman J, Hasan MK (2012) Newton-EGMSOR methods for solution of second order two-point nonlinear boundary value problems. J Math Syst Sci 2:185–190
18. Maleknejad K, Nediasl K (2011) Application of sinc-collocation method for solving a class of nonlinear fredholm integral equations. Comput Math Appl 62:3292–3303
19. Sahu PK, Ray SS (2013) Numerical approximate solutions of nonlinear fredholm integral equations of second kind using B-spline wavelets and variational iteration method. Comput Model Eng Sci 93(2):91–112
20. Nadir M, Khirani A (2016) Adapted Newton-Kantorovich methods for nonlinear integral equations. J Math Statistics 12(3):176–181
21. Saudi A, Sulaiman J (2016) Path planning simulation using harmonic potential fields through four-point-EDGSOR method via 9-Point Laplacian. Jurnal Teknologi 78(8–2):12–24
22. Akhir MKM, Othman M, Sulaiman J, Majid ZA, Suleiman M (2011) The four point-EDGMSOR iterative method for solution of 2D Helmholtz equations. Commun Comput Information Sci 253(3):218–227
23. Dahalan AA, Sulaiman J, Muthuvalu MS (2014) Performance of HSAGE method with Seikkala derivative for 2-D fuzzy poisson equation. Appl Math Sci 8(17–20):885–899

# Factors Affecting Government Employees' Acceptance of EDMS: A Systematic Review


**Bridget Geoffrey Lojonon and Rayner Alfred**

**Abstract** Archiving and storing information is a vital aspect that every government strives to optimize while serving citizens. An electronic document system ensures that the flow of information from storage, processing to transmission is as seamless as possible. The Electronic Document Management System (EDMS) is one such system that governments from around the world have been striving to implement. However, a hindrance has been observed in the acceptance of EDMS by government employees, which has been attributed to several factors. This work highlights a systematic literature review of the factors that affect the acceptance of the system. Additionally, a quantitative approach through a questionnaire is used to determine which factors affect the acceptance of EDMS among government employees. A lack of awareness was established as the primary factor affecting the acceptance of EDMS. It was concluded that the government should put more effort into ensuring that more people are aware of any systems that are implemented by the government.

**Keywords** Electronic document management system (EDMS) · Government employees · Adoption behavior · eGovernment · Technology acceptance

## 1 Introduction

Governments around the world have been incorporating Information Communication Technology (ICT) in most of their entities. The aim of utilizing ICT is to ensure better interaction with the employees and better delivery of services [1]. Since governments manage a lot of information and documents, a robust and effective data management system is a necessity [2]. For a government to successfully incorporate ICT in its activities, more effort has to be put into developing a conducive

B. G. Lojonon · R. Alfred (✉)
Knowledge Technology Research Unit, Universiti Malaysia Sabah, Jalan UMS, Kota Kinabalu, 88400 Sabah, Malaysia
e-mail: ralfred@ums.edu.my

B. G. Lojonon
e-mail: bglojonon@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_28

organizational structure, continuous data flow, and reducing costs [3, 4]. The primary goal of using ICT in the government is usually assumed to be to provide information services to the citizens. However, it is used to develop strategic links among government bodies, facilitate government transactions and ensure smooth communication between different departments within the government as well [5]. Hence, governments have been pushing the implementation of the EDMS.

This work aimed to highlight the factors that affect the acceptance of the government employees of EDMS usage and determine which factors should be prioritized. Comparatively, in terms of the adaptation and implementation of EDMS, previous studies include factors such as organizational challenges in managing the systems, user resistance and support, system integration needs, accelerating change management strategies, lessons learned as well as understanding past implementation of the systems. Similarly, this work and most of the previously published literature are central to understanding users' perception towards improving information system usage and its effects, and they aimed to address the issue of identifying all possible factors that can influence the decisions of the organizations and their stakeholders. In general, our results show that EDMS adoption and implementation involves several technological, organizational and user-related factors. The work presented contributes to the literature by identifying a list of factors considered critical to the effective implementation of EDMS, as well as determining the relative importance of these factors in the context of their priority which is a not well evaluated and even neglected area as recommended by McLeod et al. (2011).

The Electronic Document Management System (EDMS) is a system powered by ICT and utilized by the government to manage and process valuable data and information of their employees. EDMS regulates the flow of documents and maintains how business and operations within and outside an organization are carried out [6]. The data managed by EDMS is processed and stored digitally, which greatly improves the performance and productivity of public service. The field has seen new advancements to the EDMS, which facilitate better and increased capabilities as compared to the traditional document generation methods [7]. The new system is more advanced and sophisticated and utilizes computer-based information systems that processes, analyzes, saves and disseminates information at high speeds and presents the data in a comprehensible manner useful to the users and policymakers [8]. However, the acceptance of EDMS by government employees has been affected by several factors [9]. This systematic review highlights the implications of EDMS and studies the factors that influence its acceptance by government employees.

## 2 The Concept of Electronic Document Management System

Before the systematic review is carried out, understanding the concept of EDMS is essential. The process of creating, retrieving, storing, modifying, displaying and

storing data is defined as data management [10]. Before the arrival of the internet and technology, data management systems had a different concept on document management. Technology has changed and revolutionized how traditional file management systems work. Burtylev et al. (2013) and Goings et al. (2007) point out that shifting from paper-based documents to electronic filing systems has defined how documents, especially those for web content, are managed as well as the management of users responsible for the content. As a result, old content management systems were re-developed to manage and control content sources, allot administrative tasks, process transfer of files, and control workflow [11]. The implementation of the electronic management system is a result of evolution of technology. Therefore, the EDMS was developed in a bid to meet the challenges of the information revolution. While attempting to understand EDMS, understanding what a document refers to in a management system is also critical. In document management systems, a document is any information that is structured and recorded into an item that can be accessed or consumed by a person [12]. Abbasi et al. (2016) argue that a document as storage or record of information which may be in the form of recorded speech converted into transcripts. However, from a technological standpoint, a document can be regarded as a piece of information that can consist of several other types of data and can exist in different areas but within the same network [13]. The information could contain a record of data that holds specific attributes like employees' entities in an organizational or governmental setting [14]. The document is what defines an electronic management system.

## 3 Systematic Literature Review: Method

The method illustrates all the processes carried out to ensure that the review was successful. This literature review has been carried out based on the review protocol proposed by Kitchenham and Charters [15]. For this case, the study aims to identify the factors that affect the acceptance of EDMS by government employees. The main focus of the method is to determine why and how the sources were acquired for a study [16]. The following are the steps that were considered while carrying out a systematic literature review.

### *3.1 Research Questions*

A strategic literature review must have a set of research questions or objectives. These questions and objectives will guide the study [17]. The research questions (RQs) were formulated to define the scopes of the research according to three viewpoints: population, intervention and outcomes [18].The population viewpoint covers the areas or roles (e.g., EDMS adoption and implementations) affected by the intervention. In Information System (IS) management, the populations might be any of

the following: A specific IS role or a type of IS tool and its application area. Then, the intervention viewpoint covers IS technologies that address specific issues (for example, EDMS technologies to perform specific tasks such as data management). Finally, the outcomes viewpoint should relate to factors of importance to practitioners such as improved management of information flow, administrative tasks and processing transfer of files. The research questions highlighted by the study areas are outlined as follows:

1. RQ1—Which factors significantly affect the government employee's acceptance of EDMS in the public sector?
2. RQ2—Which of these factors are the most significant/important to improve EDMS acceptance among government employees where technology adoption is mandatory?

## *3.2 Defining the Literature Body*

The approach taken in the collection of relevant literature for the study consists of one part. However, an investigation could have several strategies through which relevant literature could be collected [19]. The approach used is a general library search of the research keywords in all related libraries. The keywords were run through several libraries to find the relevant literature. The main keywords used in the search include "EDMS" and "e-Government". Using keywords when searching for literature in libraries provides a specific range through which the documents will be obtained [20]. The two keywords provided exclusive results which excluded other types of document systems and organizations, respectively. Since the purpose of the study was to identify the factors that affect the acceptance of EDMS by government employees, no other types of management systems are discussed. Additionally, the focus is only on government employees and not those from other organizations. Therefore, this systematic review does not include studies on different types of systems or employees from other organizations other than the government.

However, the selected literature body from the search libraries had to adhere to specific inclusion criteria. First, the literature used should address electronic document management systems in the government or e-government as the primary or secondary area of the research. Therefore, the keywords "EDMS" or "e-government" should be present predominantly in the text. Second, the selected literature has to be research papers or peer-reviewed journals. Third, the text used has to be written in English; no other languages are acceptable. And last, the literature used should have a document body longer than one page. Additionally, the literature used should not contain books, keynotes, presentation notes or extended abstracts. However, the literature body was obtained through the following steps:

1. Collection of the literature. Collecting literature involved the search through relevant libraries using the keywords. The use of keywords ensures that a maximum number of research papers are found.

2. Application of inclusion/exclusion criteria. The previously discussed inclusion criteria for the research papers to be used is observed.
3. Verification of rejected papers. If an article fails to meet the inclusion criteria, they are still verified to check whether some of the keywords exist in the document. Therefore, it would eliminate the cases where papers are rejected but still are connected to the initial research ideas.
4. Verification of accepted/included papers. Once the papers pass the inclusion criteria, it is checked manually by reading the conclusion as well as the abstract. In this particular step, a research paper is checked whether the research papers used in the review address the factors that affect the acceptance of EDRM among government employees.

### *3.3 Collection of the Literature Body*

The systematic literature review protocol by Kitchenham was used to obtain the literature body for this review. The extraction date for the information was around May 2020. The four steps for defining literature as described above are the ones usually utilized in the study [21]. The collection of literature started with a library search. The number of papers initially found were 22. This includes all the research papers with the keywords of the study. Once the inclusion/exclusion criteria were applied to the literature, 11 articles remained, as 11 of them were rejected. Once the second step of using the inclusion/exclusion criteria was through, the third step was carried out, but no rejected papers were included in the literature. Finally, the fourth step of verifying the accepted literature resulted in four more documents being rejected. The final remaining number of documents after the fourth step was 11.

The first step of collecting literature produced 22 papers because little research has been done on the factors that affect the acceptance of EDMS by government employees. Additionally, the search in the libraries brought articles that only mentioned the EDMS or the use of the system in e-Government. The second step, which applies the inclusion/exclusion criteria, rejected seven papers because they lightly highlighted the research topic and had little research on the factors that affect the implementation of EDMS among government employees. The third step failed to reject any papers like all of the rejected documents was unable to meet the requirements of the literature needed for this review. The four documents dismissed in the last step of collecting literature, after a thorough reading of their introductions and conclusions, failed to indicate the factors that affect acceptance of EDMS but stated generally why implementation of the system is hard. During the literature collection process, the papers are read to extract information that would address the research questions and objectives [22]. Table 1 summarizes the literature collection process and the number of papers at each step.

**Table 1** Literature collection

| Step | Number of papers |
|---|---|
| • Collection of literature | 22 |
| • Application of inclusion/exclusion criteria | 11 |
| • Verification of rejected papers (included papers) | 0 |
| • Verification of included papers | 11 |

### *3.4 Analysis*

The analysis section provides an insight into the literature used, as well as the findings of the systematic review. Besides the research, a questionnaire is prepared and presented to government employees to determine the factors that hindered them from accepting the use of EDMS in the government. The results of the survey are analyzed in this section. This section addresses the first research question and objective. The literature covers the factors that affect the adoption of EDMS by government employees. Thus, the first research question is addressed. Since the second research question and purpose aim at developing a model, it would be addressed after the following analysis.

## 4 Factors Affecting Implementation of EDMS

During the systematic literature review, a general overview was formed where the factors that affect government employees' acceptance of EDMS are identified, synthesized and categorized. Although the aim was to determine the factors that specifically influence government employees, most of the existing literature provide general factors. Implementation of EDMS is a complicated and vast issue that involves several organizational, technical or technological, and user-related factors [23]. The common factors for EDMS implementation according to organizational factors include collaboration, legislation environment, strategic planning, budget cost, and top management support; technical factors include system integration, data quality, user requirements, security and privacy trust, IT implementation team and ICT infrastructure; and user factors include resistance to change, staff training and awareness. The study results generated 40 critical success factors that influenced the use of EDMS among government employees from 15 articles. These critical success factors are categorized under four main variables of the UTAUT framework namely, "Performance Expectancy", "Effort Expectancy", "Social Influence", and "Facilitating Condition".

Meanwhile, based on the study results, organizational factors related to the implementation of EDMS can be associated with rules, procedures and processes outlined by the organization, and for this particular case, the government [24]. Technical

problems that may be encountered while implementing EDMS may include software design, quality and security of data [25]. The user-related factors are related to resistance to change, training and awareness [26], which will be analyzed further since the objective of the study is to determine factors affecting the employee's acceptance of the system.

## *4.1 Resistance to Change*

Change is not always well-perceived by an organization, a person or a group of people. Most of the literature on the study have indicated that resistance to change is a common factor that affects an employee's acceptance of EDMS. A person or a group of individuals may resist change if they believe that it may be of risk or unsuitable for them. According to Maguire (2010), resistance is considered any employee action that aims at challenging, disrupting and inverting prevailing assumptions to power. Resistance to change is one of the primary challenges during the implementation of EDMS [23]. The processes involved in implementing EDMS bring a lot of changes to all levels of an organization's departments and divisions. Additionally, the implementation may result in a change in tasks carried out by the employees or even switch in leadership. According to [27] if the change is not monitored, it will grow gradually to unprecedented levels. Therefore, resistance to the implementation of EDMS is expected.

## *4.2 Training*

When a new ICT system is being implemented, training has to be undertaken. Training the staff on new Information Technology (IT) skills while adopting a new system plays a significant role in its implementation. Most of the failures of implementation of EDMS have significantly been affected by lack of sufficient training of the employees [28]. A new ICT system may require advanced technical skills which may be difficult for some of the employees to master. The success of EDMS is highly attributed to the skill training that an individual possesses. Additionally, Leikums (2012) and McLeod et al. (2011) indicated that training courses for the employees have to be considered to facilitate their awareness in dealing with the EDMS positively. Therefore, training is another factor that has affected employee acceptance of EDMS.

### *4.3 Awareness*

Before any change is implemented in an organization or the government, the staff have to be informed first. While implementing a new system, the employees have to be made aware of the enhancement that the new technology would bring to their normal working process [28, 29]. For instance, while implementing EDMS, it is essential to cite the significance of the system to the pre-existing record keeping practices. Thus, employees would be aware of the new system and what it entails. According to Yaacob and Mapong (2011) and Asogwa (2012), the success of EDMS will depend highly on the awareness made on the program. The implementation of the system would be a failure if the people involved are not made aware. Akhavan et al. (2006) state that awareness is directly dependent on support from the top management. Awareness is a crucial factor in ensuring that the implementation of EDMS is a success.

## 5 Discussion and Limitations

This section elaborates the factors affecting implementation of EDMS by summarizing and discussing the findings in relation to RQ1—Which factors significantly affect the government employee's acceptance of EDMS in the public sector (e.g., organizational, technological or user-related)? Highlighting the factors affecting the acceptance of the government employee's towards EDMS usage contributes to the recognition of the predictors that will determine the full realization of EDMS programs. Table 2 tabulates all the factors related to EDMS acceptance by government employees.

In relation to RQ2—Which of these factors are the most significant/important to improve EDMS acceptance among government employees where technology adoption is mandatory? Based on the type of issue addressed, the task of prioritizing the factors to find the relative importance to improve EDMS acceptance among government employees is using the Analytic Hierarchy Process (AHP) methods. Based on the results, the findings reveal that "Performance Expectancy" and "Effort Expectancy" are the two most significant/important factors which affect the use of EDMS among government employees.

During the process of SLR, we noticed 14 issues that influence the implementation of EDMS. These factors are believed to be common when it comes to implementing EDMS in the government. Implementing EDMS is complex and entails user, technical and organizational related factors. For example, Artamonov et al. (2018) noted that implementing EDMS can be adversely affected by the lack of basic legislative regulations on the national level as well as within the organizational framework. Technical factors describe data quality, security and software design. For example, Aziz et al. (2018) clarified that poor ERM can result in costly legal liabilities. EDMS should be safe from misuse or undocumented alteration. User factors entail issues

**Table 2** Critical success factors and definition of the UTAUT

| UTAUT factors | Critical success factor | Works |
|---|---|---|
| Performance Expectancy | • EDMS functionality;<br>• Usability and understanding of output;<br>• Demonstration of benefits;<br>• Piloting and testing;<br>• Integration of systems and technology;<br>• User friendliness of EDMS;<br>• Efficiency of EDMS;<br>• Effectiveness of EDMS | [3, 4, 12, 21, 30–34] |
| Effort Expectancy | • Process readiness;<br>• Infrastructure readiness;<br>• Architecture readiness;<br>• Support of team-based approaches to problem solving;<br>• Spirit of cooperation and teamwork;<br>• Sharing of expertise;<br>• Change of management;<br>• Assurance that a project has a clear agenda;<br>• Alignment of projects with business objectives;<br>• Communication;<br>• Policies and guidelines | [3, 30, 31, 35, 36] |
| Social Influence | • Management, leadership, and commitment toward EDMS;<br>• Colleagues' recommendation of the use of EDMS;<br>• Top management's recommendation of the use of EDMS;<br>• Subordinates' support of the use of EDMS;<br>• Impact of EDMS on reputation;<br>• High regard of people using EDMS;<br>• Planning and project management;<br>• Gaining of commitment and support of chief executive officers;<br>• Top management encouragement toward informal or formal communication;<br>• Development of clear mission regarding business objectives;<br>• Top management encouragement toward use of EDMS | [30, 31, 37, 38] |

(continued)

**Table 2** (continued)

| UTAUT factors | Critical success factor | Works |
|---|---|---|
| Facilitating Conditions | • Active encouragement of employee participation in EDMS-related decisions;<br>• Involvement of all levels within the organization and external stakeholders;<br>• Consistent updates of management knowledge;<br>• Involvement of EDMS end users;<br>• Adequate training and support for users;<br>• Requirement-driven procurement planning;<br>• Provision of technical resources (e.g., equipment, software);<br>• Human resource availability;<br>• Sufficient monetary resources provided to support EDMS implementation;<br>• Prior development or existence of necessary infrastructures | [4, 6, 21, 30, 31, 36, 39] |

related to resistance to change, drivers, training and culture. For example, employees' resistance to change is a challenge in implementing information systems. Governments should seek to make more of their employees aware of EDMS implementation to increase its efficiency.

### *5.1 Limitations*

Like any other research, this study has its limitations. The study derived its variables from a systematic review which depicts limitations. Additionally, the model was not designed to include a reliability and validity test, which should be included in future research. Similarly, this work finds that previous empirical and theoretical studies have described the technical aspects of EDMS, but limited studies have evaluated the components of EDMS [4, 40]. Limitations faced by other studies included the need to conceptually refine and empirically test both model and indicators, dependency on quantitative data, in-depth review on DMS integration, limited number of case study, the need for more research data and the scope of study to include institutional units.

## 6 Unified Theory of Acceptance and Use of Technology (UTAUT)

Understanding the different contextual factors that influence effective adoption of EDMS requires a careful grasp of the different frameworks and approaches used in different contexts. As stated by Davis et.al. (1989), several theories and models have been used to provide a theoretical base for examining factors that influence technology adoption in organizations [58]. The Unified Theory of Acceptance and Use of Technology (UTAUT) outline is a variation of another acceptance model, the Technology Acceptance Model (TAM), that was initially implemented to be applied in new systems. The UTAUT model addressed the essential principles of various conceptual models. The UTAUT model has four variables. However, of all the variables, the most outstanding is performance expectancy. The UTAUT model has improved the prediction of technology usage by more than 65% and is, therefore, a suitable framework [32]. As compared to TAM, the UTAUT model is superior because the former can only predict up to 30%. Other acceptance models have an acceptance ratio ranging between 17 and 50% [41]. The UTAUT model framework has been applied in different countries globally and has proved to be effective.

The UTAUT model has applied and employed several aspects since its inception. Some of the aspects adopted by the model include effort expectancy, facilitating conditions, social influence and performance expectancy [42]. The extent to which an individual can contemplate integrating technology to improve their

work performance is referred to as performance expectancy [43]. In the case of this systematic literature review, the performance expectancy is the expectation government employees will have on the enhancement of performance if the new document management system is implemented. Performance expectancy is a primary aspect that is considered by individuals when trying to accept or use a particular technological system [34]. However, this study has demonstrated that awareness is one of the critical factors that must be considered before the EDMS is implemented. Only after the employees are aware of the system can they have performance expectancy.

Although the literature collected was inconsistent in outlining factors that affected the acceptance of EDMS by government employees, the UTAUT model could be used as a confirmation tool. The following hypotheses were developed based on the UTAUT model. Additionally, the hypotheses were consistent with the projections of the research.

- **Hypothesis 1**: Effort expectancy negatively relates to behavioral intention.
- **Hypothesis 2**: Social factors positively relate to behavioral intention.
- **Hypothesis 3**: Performance expectancy negatively relates to behavioral intention.
- **Hypothesis 4**: Attitude towards acceptance of a new management system is positively related to behavioral intention.
- **Hypothesis 5**: Effort expectancy negatively relates to attitude towards technology.
- **Hypothesis 6**: Social factors are positively related to attitude.
- **Hypothesis 7**: Attitude is positively related to performance expectancy.
- **Hypothesis 8**: Facilitating conditions positively relate to behavioral intention.

## *6.1 Questionnaire and Analysis*

A questionnaire was chosen as the appropriate means of collecting data for this study because of several reasons. First, it is easier to construct and administer. Second, once the responses to the questionnaire are obtained, they are easy to analyze. Additionally, most people are aware of and familiar with surveys. Therefore, it is very suitable for research. Last, the questions in the questionnaire are straightforward, and the results are usually relatively correct. According to Lopatovska and Arapakis (2011), questionnaires are the best means of collecting data of a relatively larger population. Other than indicating the suitability of the data collection method, the authors also use the approach in their research, which is related to Information System (IS). The questionnaire analysis covers two significant areas: descriptive and quantitative studies.

### *6.2 Factors Affecting the Acceptance of EDMS*

The secondary analysis in the questionnaire attempts to determine the factors that affect the acceptance of EDMS by government officials. This section includes questions related to demography profile (gender and age) and the use of IT in their field of work.

**Computer Usage at Work**:

Does your work involve the use of a computer? (Respond: Yes or No).

**Document Handling in Office**:

Do you use your computer or hard copy papers to handle documents? (Respond: Computer or Hard copy Paper).

**Awareness of Implementation of EDMS**:

Were you aware the government was implementing EDMS? (Respond: Yes or No).

**Training on the EDMS**:

Were you given enough training on the EDMS? (Respond: Yes or No).

**Acceptance of the EDMS**:

Do you accept and approve the adoption of EDMS? (Respond: Yes or No).

## 7 Conclusion

In use, more often than not, EDMS users and the system itself are in a conflicting position. The results do not mention that EDMS is poorly designed or that users are underprepared. It is natural that the opposing relations may exist among the government employees. It can only be speculated that the reason for this is the highly complex nature of EDMS. EDMS is more than a simple technological system, and the evaluation of such a complex system is difficult. The government should understand that its employees hold the key to success of adopting and implementing EDMS.

## References


1. Lee K, Choi SO, Kim J, Jung M (2018) A Study on the factors affecting decrease in the government corruption and mediating effects of the development of ICT and E-Government—a cross-country analysis. J Open Innov Technol Market Complexity 4(3):41
2. Tagbotor DP, Adzido RYN, Agbanu PG (2015) Analysis of records management and organizational performance. Int J Acad Res Accounting Finance Manage Sci 5(2):1–16
3. Radzi MA, Yatin SF, Fadzil NA, Aziz SA (2018) Document preparation for electronic document management system. Int J Acad Res Business Soc Sci 8(9):179–190
4. Karlos AN, Nengomasha CT (2018) Change management: a critical factor for successful implementation of an electronic document and records management system (EDRMS): a Namibian case study. UNAM. Retrieved from https://repository.unam.edu.na/handle/11070/2421
5. Bunawan AA, Nordin S (2015) The challenges in preserving the electronic records metadata. Int J Information Syst Eng 1(1):1–7
6. Alshibly H, Chiong R, Bao Y (2016) Investigating the critical success factors for implementing electronic document management systems in governments: evidence from Jordan. Information Syst Manage 33(4):287–301
7. Viau JH (2015) Employee records: what to keep, how to keep, and when to shred. J Med Pract Manage MPM 30(4):258
8. Wohlin C (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering, pp 1–10
9. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Bouwman J (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3
10. Abbasi A, Sarker S, Chiang RH (2016) Big data research in information systems: toward an inclusive research agenda. J Assoc Information Syst 17(2):3
11. McLeod J, Childs S, Hardiman R (2011) Accelerating positive change in electronic records management: headline findings from a major research project. Arch Manuscripts 39(2):66–94
12. Artamonov A, Ionkina K, Tretyakov E, Timofeev A (2018) Electronic document processing operating map development for the implementation of the data management system in a scientific organization. Proc Comput Sci 145:248–253
13. Mosweu O, Bwalya KJ, Mutshewa A (2016) A probe into the factors for adoption and usage of electronic document and records management systems in the Botswana context. Information Dev 33(1):97–110. https://doi.org/10.1177/0266666916640593
14. Guetterman T (2015) Descriptions of sampling practices within five approaches to qualitative research in education and the health sciences
15. Achimugu P, Selamat A, Ibrahim R, Mahrin MNR (2014) A systematic literature review of software requirements prioritization research. Inf Softw Technol 56(6):568–585
16. Boell SK, Cecez-Kecmanovic D (2014) A hermeneutic approach for conducting literature reviews and literature searches. Commun Assoc Information Syst 34(1):12
17. Su T, Wang SX, Chen Y, Tsou T, Cheng J (2017) Investigating the usability of electronic document management systems in government organizations from a human factor engineering perspective. J Adv Manage Sci 5:14–17
18. Mosweu O, Bwalya K, Mutshewa A (2016) Examining factors affecting the adoption and usage of document workflow management system (DWMS) using the UTAUT model. Rec Manage J 26(1):38–67
19. Bozorgkhou N (2015) An internet shopping user adoption model using an integrated TTF and UTAUT: evidence from Iranian consumers. Manage Sci Lett 5:199–204
20. Kim SH (2014) A study on adoption factors of Korean smartphone users: a focus on TAM (Technology Acceptance Model) and UTAUT (Unified Theory of Acceptance and Use of Technology). Adv Sci Technol Lett 57(1):27–30

21. Rosa AT, Pustokhina IV, Lydia EL, Shankar K, Huda M (2019) Concept of Electronic Document Management System (EDMS) as an efficient tool for storing document. J Critical Rev 6(5):85–90
22. Leikums T (2012) Managing human factors in implementing electronic document system in the public sector. Challenges of the Knowledge Society, 2046
23. Yaacob RA, Mapong Sabai R (2011) Electronic records management in Malaysia: a case study in one government agency, pp 420–433
24. Leikums T (2012) A study on electronic document management system integration needs in the public sector. Int J Adv Eng Technol 5(1):194–205
25. Akhavan P, Jafari M, Fathian M (2006) Critical success factors of knowledge management systems: a multi-case analysis. Eur Business Rev 18(2):97–113
26. Grange M, Scott M (2010) An investigation into the effect of poor end-user involvement on electronic document management system (EDMS) implementation
27. Asogwa BE (2012) The challenge of managing electronic records in developing countries. Rec Manage J 22(3):198–211
28. Nguyen LT, Swatman P, Fraunholz B, Salzman S (2009) EDRMS implementation in the Australian public sector. In: ACIS 2009: evolving boundaries and new frontiers: defining the IS discipline: Proceedings of the 20th Australasian conference on information systems, ACIS, pp 915–928
29. Goings DA, Johnson JJ, Marshall B, Goette T (2007) The influence of government regulations on content management systems: an exploratory study. Commun IIMA 7(1):65–76
30. Aziz AA, Yusof ZM, Mokhtar UA, Jambari DI (2019) The intention to adopt electronic document and records management system: questionnaire development procedure. Int Conf Electrical Eng Informatics (ICEEI) 2019:590–595
31. Aziz AA, Yusof ZM, Mokhtar UA (2019) Electronic Document and Records Management System (EDRMS) adoption in public sector-instrument's content validation using content validation ratio (CVR). J Phys Conf Ser 1196, 1–7. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1196/1/012057/pdf
32. Putra DA, Jasmi KA, Basiron B, Huda M, Maseleno A, Shankar K, Aminudin N (2018) Tactical steps for e-government development. Int J Pure Appl Math 119(15):2251–2258
33. Kassab MI, Abu Naser SS, Al Shobaki MJ (2017) The impact of the availability of technological infrastructure on the success of the electronic document management system of the Palestinian Pension Authority. Int J Eng Information Syst (IJEAIS) 1(5):93–109
34. Paramonova IE (2016) Electronic document-management systems: a classification and new opportunities for a scientific-technical library. Sci Techn Information Process 43(3):136–143
35. Kassab MI, Naser SSA, Al Shobaki MJ (2019) The role of policies and procedures for the electronic document management system in the success of the electronic document management system in the Palestinian Pension Agency. Int J Acad Multidisc Res (IJAMR) 3(1):43–57
36. Nengomasha C, Chikomba A (2018) Status of EDRMS implementation in the public sector in Namibia and Zimbabwe. Rec Manage J 28:252–264
37. Mukred M, Yusof ZM, Alotaibi FM, Mokhtar UA, Fauzi F (2019) The key factors in adopting an Electronic Records Management System (ERMS) in the educational sector: aUTAUT-based framework. IEEE Access 7:35963–35980
38. Ambira CM, Kemoni HN, Ngulube P (2019) A framework for electronic records management in support of e-government in Kenya. Rec Manage J https://doi.org/10.1108/rmj-03-2018-0006
39. Abidin SSZ, Husin MH (2018) Improving accessibility and security on document management system: a Malaysian case study. Appl Comput Informatics. https://doi.org/10.1016/j.aci.2018.04.002
40. Madigan R, Louw T, Dziennus M, Graindorge T, Ortega E, Graindorge M, Merat N (2016) Acceptance of automated road transport systems (ARTS): an adaptation of the UTAUT model. Transp Res Proc 14:2217–2226
41. Burtylev IN, Mokhun KV, Bodnya YV, Yukhnevich DN (2013) Development of electronic document management systems: advantage and efficiency. Sci Technol 3(2A):1–9

42. Bannister F, Connolly R (2014) ICT, public values and transformative government: a framework and programme for research. Government Information Q 31(1):119–128
43. Thompson N, Ravindran R, Nicosia S (2015) Government data does not mean data governance: lessons learned from a public sector application audit. Government Information Q 32(3):1–7

# Prioritization of Factors Affecting Government Employees' Acceptance of EDMS Using the Analytic Hierarchy Process (AHP) Method


**Bridget Geoffrey Lojonon and Rayner Alfred**

**Abstract** Document management (DM) is fundamentally one of the most effective approaches applied in managing information flow in an organization. In reality, the utilization of an electronic document management system (EDMS) is depicted essentially when documents are used as memory storage for companies and record-keeping portals that document how operations are achieved. Consequently, the study aimed to respond to the recommendation of McLeod et al. (Arch Manuscr 39:66–94, 2011 [5]) to explore the predictors of EDMS implementation to advise on planning the successful use of EDMS programs. The first objective was to develop a new theoretical model based on the unified theory of acceptance and use of technology (UTAUT) to study the user acceptance and adoption of the EDMS among government employees. The second objective was to determine the factors affecting EDMS acceptance and adoption among government employees that should be given priority using the analytic hierarchy process (AHP) methods. The third objective was to determine whether the constructs of UTAUT influence behavioral intention to use EDMS among government employees. The research employed a systematic review that was developed to identify different factors affecting government employees' acceptance and adoption of EDMS. The critical success factors were generated based on four levels of UTAUT from 15 articles. The study's results, from the systematic review, yielded 40 factors that influenced the use of EDMS among government employees.

**Keywords** Electronic document management system (EDMS) · Government employees · Adoption behavior · E-government · Analytic hierarchy process (AHP) · Prioritization

B. Geoffrey Lojonon · R. Alfred (✉)
Knowledge Technology Research Unit, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
e-mail: ralfred@ums.edu.my

B. Geoffrey Lojonon
e-mail: bglojonon@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_29

## 1 Introduction

Many organizations understand that the effective regulation of information flow results in effective management. Document management (DM) has been increasingly realized as one of the most effective approaches applied in managing information flow in an organization in both the private and public sectors [1, 2]. Essentially, the documents are used as memory storage of companies and record the way operations are achieved. DM allows the easy retrieval, identification, and management of information. The main role of an electronic document management system (EDMS) is to enable the free flow of records through the institutions as well as ensure that information is available when requested [1]. EDMSs have increasingly become critical to governmental organizations. Although the implementation cost of such a system is high, an EDMS allows an institution to exemplify effective performance through reducing costs, improving capacity, minimizing errors, and saving on labor [3]. The implementation of an EDMS is a critical factor in forming a virtual workplace setting and transforming the dynamics of modern companies and their staff.

While EDMSs are significantly becoming an important component of infrastructure in many organizations, their uses are not well understood because of limited studies on predictors for the effective implementation of EDMS [4, 5]. While most empirical and theoretical studies have described the technical aspects of EDMSs, limited studies have evaluated their components. Therefore, there exists a need to explore the predictors for EDMS implementation to recognize the factors that will determine the full realization of EDMS programs.

Previous studies on the user acceptance of DM systems that aimed to identify, investigate, analyze, and propose various models of technology adoption adopted different methods to prioritize the most desirable factors affecting EDMS projects. However, hardly any studies have been found to specifically use the analytic hierarchy process (AHP) method to prioritize the factors affecting government employees' acceptance of EDMS. For example, the analytical descriptive method was adopted to identify senior management commitment toward EDMS development and support, the impact of technological infrastructure on EDMS success, and the role of policies and procedures for EDMSs as well as to analyze the impact of top management support on EDMSs [6–9]. A phenomenological design was used to investigate the effects of DM on the implementation of an e-government [10]. Interviews and questionnaires were used to investigate the critical factors of the EDMS and to identify its security and accessibility issues, its level of acceptance, and the influence of individual, technological, and environmental factors on its adoption [11–13]. An experiment-based method was used to investigate the usability of EDMSs from a human factor engineering perspective [14]. A comprehensive systematic review of the relevant literature was made to critically explore issues and practical strategies to support accelerating the pace of positive change in managing electronic records [5].

An investigation into the predictors of EDMS implementation will reveal valuable information to plan the successful use of EDMS programs. The research responds to

the recommendation of McLeod et al. [5] of a study in this field. The research seeks to determine the predictors of government employees' acceptance and adoption of EDMS. The study adopts the unified theory of acceptance and use of technology (UTAUT) as a foundation of exploring the predictors for EDMS use. The outcome of this research will help government agencies understand the critical factors to effectively implement EDMSs. Furthermore, the research identifies the relative importance of these predictors in the context of their priorities, which has been commonly ignored in the literature. The rest of the research is structured as follows. In the first section, the development of the theoretical underpinning of the research is shown. The research then describes the search strategy applied in the development of the literature review. Finally, the outcome of the literature review is presented using tables.

## 2 The Unified Theory of Acceptance and Use of Technology

As stated by Davis et al. (1989), several theories and models have been used to provide a theoretical base for examining factors that influence technology adoption in organizations [15]. The UTAUT framework is an all-encompassing version of the technology acceptance model (TAM), which was designed for a new system use context. The UTAUT model addressed the fundamental principles of different conceptual models, comprising innovation diffusion theory (IDT), social cognitive theory (SCT), TAM, the theory of reasoned action (TRA), the theory of planned behavior (TPB), and the motivational model. The UTAUT derives 32 variables from these theories to downscale them into four variables, namely, social influence, facilitating conditions, performance expectancy, and effort expectancy. The research considered the UTAUT to be a suitable framework as it can improve the predictors of technological use to as much as 70% [3]. This has been demonstrated to be superior to the TAM, whose prediction accuracy can only reach up to 30%, while those of other models range from 17 to 54%. Furthermore, international organizations have utilized the UTAUT to predict technological adoption in Yemen, Tanzania, Botswana, Turkey, and New Zealand [2, 3]. Therefore, the selection of the UTAUT framework appears justifiable in this study.

Different factors are adopted and applied from the UTAUT model. Such elements comprise social influence, performance expectancy, facilitating conditions, and effort expectancy. Studies describe performance expectancy as the level to which people consider using technology to help achieve job performance. It reflects the government employees' perception of expected performance improvement when using an EDMS in handling the organization's DM [16]. Studies demonstrate that performance expectancy is the main factor for users to accept, use, or depict the intention to use specific technology [17–19]. This factor is in line with the perceived usefulness of the EDMS in the TAM, and it describes people's belief that using a specific system will allow them to contribute and enhance their task performance. Several items can be used to measure performance expectancy, including fast responses in

the likeliness of any changes when using the EDMS, the improvement of task accuracy when using the EDMS, the reduction of work task handling times when using the EDMS, the reduction in decision-making times when using the EDMS, and the extensive management of life-cycle information when using the EDMS.

The literature describes effort expectancy as a level of ease related to the use of technology. Effort expectancy demonstrates the complexity of adopting the EDMS in the tasks of government employees. When government workers perceive that the EDMS demands minimal effort and is easy to use, high acceptance toward the EDMS is expected. This variable corresponds to the EDMS's perceived ease of use in the TAM. Effort expectancy is also one of the primary attributes of the UTAUT to predict usage, adoption, or behavioral intent to utilize a technology [17, 18]. Howard et al. [17] described three scale items to measure effort expectancy: the EDMS guideline is easy to follow; information among the stakeholders is easily exchanged following the adoption of the EDMS; and cooperation with the EDMS can be easily learned.

Studies describe social influence as the level to which people believe that others should use the technology. Social influence is in line with the subjective norm of the TAM. The variable signifies the effect of external attributes such as the opinions of friends and relatives on the government employees' behavior. Their opinion will affect the usage and adoption of the EDMS. Afshan [20] described the subjective norm as individuals' opinions that most people who influence them think they should perform or not perform certain behaviors in context. They also demonstrated how social influence is a critical element for technology usage, adoption, and intention. The concept of social influence applies to real-life situations as it describes the type of prestige that society is encouraged to follow. Bozorgkhou [19] noted two levels of social influence. The first level is technological use, where family and friends in other organizations using certain technologies encourage people to do the same. The second level entails the outcome of using such technologies. Specifically, users expect excellent feedback from the social setting following the achievement of high performance because of new technological use. Two scales that measure social influence include support from the organizational unit to use EDMSs and people using EDMSs because of a high percentage of coworkers who use EDMSs.

Studies describe facilitating conditions as the level to which people believe that governmental infrastructure supports their use of technology. The EDMS, as a new form of technology, demands that the government employee acquire new skills, such as computer skills and EDMS management. The government employee will not use the EDMS if they do not have the necessary operational and management skills. The literature demonstrates the relevance of facilitating conditions as the main attributes of the UTAUT framework to predict user adoption [17, 18, 20]. Three scales measure facilitating conditions: the availability of certain people for assistance with EDMS complexities; the existence of specialized instructions for the EDMS; and guidance for the selection of EDMS tools.

## 3 Critical Success Factors from the Literature

The research focused on previous studies in EDMS. In the literature review, the sources included the following: LIS databases, computer science databases, other general databases (e.g., Web of Science), and Scopus©. Both quantitative and qualitative studies were included for consideration in this review. The articles were included if (1) they measured the predictors of EDMS acceptance, (2) they included public or government employees, and (3) they were published from 2016 to 2020. Based on.

Table 1, common factors for EDMS implementation can be classified as follows: organizational factors (i.e., collaboration, legislation environment, strategic planning, budget cost, top management support); technical factors (i.e., system integration, data quality, user requirements, security and privacy trust, IT implementation team, ICT infrastructure); and user-related factors (i.e., resistance to change, staff training, awareness). Table 2 identifies the relative importance of these factors in the context of their priorities. Based on the abovementioned reviews, 40 factors have been identified and categorized into four main variables of the UTAUT framework: "Performance Expectancy (PE)," the level to which the government employee believes that the EDMS can help improve job performance; "Effort Expectancy (EE)," the level of ease related to EDMS usefulness; "Social Influence (SI)," related to government employees who can be affected by the behaviors and attitudes of other people and vice versa; and "Facilitating Conditions (FC)," the implications of governmental infrastructure in supporting EDMS use, such as resources (training), knowledge, and the user's ability. As the study aimed to respond to the recommendation of McLeod et al. [5] to explore the predictors of EDMS implementation, the research objective was to determine which system acceptance factors are the most significant/important to improve EDMS adoption and implementation. The AHP methods were used to achieve this objective by combining all the factors into a hierarchical model and quantitatively measuring their importance through pairwise comparisons [21]. During this process, the aspects of the problem, from general to detailed, were explored and then expressed in the multilevel manner required by the AHP.

Meanwhile, the key AHP methodology steps are as follows [28]:

Level 1. Developing the hierarchical (goal, criteria, and alternatives) structure of the decision problem

Level 2. Using Saaty's point scale (1–9) for the pairwise comparisons matrix for the decision problem to assess the relative weights of the criteria

Level 3. Assessing the alternative relative priority with respect to the criteria and finally calculating the overall priorities.

**Table 1** Critical success factors from the literature

| Work | Objectives | Origin of Factors | Outcomes |
|---|---|---|---|
| [22] | To evaluate the lessons learned from EDMS implementation | A case study of the Ministry of Education and Science of the Russian Federation | – Demonstration of benefits<br>– Determination of all processing materials in the scientific organization via the EDMS<br>– EDMS functionality<br>– Usability and understanding of output |
| [3] | To explore factors that affect the use and non-use of the EDMS | A case study of two public and private organizations in Nigeria | – Senior management support<br>– Culture of sharing information<br>– Providing workers with adequate information on EDMS-related principles through training |
| [10] | To examine how the use of the EDMS undermines or facilitates the application of an e-government, with the objective of recommending the best model for EDMS management in support of the e-government | An interpretive research approach of 52 participants retrieved from e-government service areas and the Kenya ICT Authority | – The top management had little to anchor the EDMS within the e-government big picture<br>– EDMS is an essential driver of the e-government |
| [6] | To determine the role of procedures and policies for EDMS implementation in the Palestinian Pension Agency | An analytical description of policies and procedures at the palestinian pension agency | – Legal and legislative guidelines for the EDMS<br>– Strategy retrieval and the backup of data in the event of an emergency<br>– Mechanism for preserving and indexing electronic documents<br>– Defining policies to generate information files for the documents |

(continued)

**Table 1** (continued)

| Work | Objectives | Origin of Factors | Outcomes |
|---|---|---|---|
| [8] | To determine the effect of technological infrastructure on EDMS implementation | A case study of the palestinian pension authority | – The existence of technological infrastructure results in the success of EDMS implementation<br>– The application of computer hardware and software results in effective EDMS implementation<br>– The application of databases and software ensures the success of EDMS implementation<br>– Applying the existing technology in the context of communication and networks results in successful EDMS implementation |
| [11] | To determine the effect of technological, individual, and environmental factors on EDMS implementation | A survey among 364 participants in higher professional education | – Security assurance<br>Policy guidance<br>– Top management support in EDMS implementation<br>– Getting the file plan right<br>– EDMSs are effective in achieving organizational objectives<br>– EDMSs are easy to learn |
| [7] | To investigate the commitment of the top management in supporting and developing EDMSs and the success of EDMSs | A case study among 43 employees in the West Bank and 65 employees in the gaza strip in Palestine | – The senior management presented a legal frame within the organizations to work through the EDMS, such as saving and signing<br>– The top management was committed to implementing policies and plans related to the EDMS<br>– The top management encouraged online work and lacked confidence in manual tasks<br>– The managers had knowledge of the nature of the EDMS<br>– The top management represented the environment of EDMS use |

(continued)

**Table 1** (continued)

| Work | Objectives | Origin of Factors | Outcomes |
|---|---|---|---|
| [14] | To examine the adoption of the EDMS in governmental institutions from a human factor engineering point of view | A survey of staff satisfaction with a human–machine interface | – System usability<br>– System function<br>– System interface |
| [12] | To examine the current EDMS in the Land Office of Malaysia | A case study of the land and district office North of Wellesley, Penang | – The EDMS is safe to use<br>– The EDMS is more efficient than a manual system<br>– The EDMS is easier to use than the manual system<br>– The employees are educated and young and do not have problems with EDMS adaptation |
| [23] | To investigate factors that influence the adoption of EDMS | A qualitative research of literature related to EDMSs | – The level to which an employee believes that the EDMS can help improve job performance<br>– The level of ease related to EDMS use<br>– The level to which an employee can be influenced by the behaviors and attitudes of other people and vice versa<br>– The implication of technical and organizational infrastructure in supporting EDMS use, such as training, resources, knowledge, and ability<br>– Security assurance<br>– Policy guidance |

(continued)

**Table 1** (continued)

| Work | Objectives | Origin of Factors | Outcomes |
|---|---|---|---|
| [23] | To explore the use of EDMSs in public service in Zimbabwe and Namibia, with the objective of highlighting the enablers and best practices each country could adopt from the other | A comparative case study of Namibia and Zimbabwe | – The process of organizational change during EDMS use<br>– Top management support in EDMS implementation<br>– The extent to which the government employee accepts and adopts the EDMS<br>– Composition of the project team<br>– Getting the file plan right<br>– Employee training on the system<br>– Resources for the ongoing support of the system |
| [24] | To analyze the use of the EDMS in limiting the challenge of low consumption levels through the lens of the information system success model and the UTAUT | Content validation by panel experts and feedback analysis using a content validity ratio | – The EDMS enables the completion of routine tasks easily<br>– The EDMS improves work performance<br>– EDMSs are effective in achieving organizational objectives<br>– EDMSs are easy to learn and easy to control<br>– The EDMS avails a user-friendly system interface<br>– Colleagues recommend the use of the EDMS<br>– Top management recommends the use of the EDMS<br>– Subordinates support the use of the EDMS<br>– The EDMS has an impact on reputation<br>– People using EDMSs are highly regarded<br>– Senior management provides good support of EDMSs<br>– The EDMS can be integrated into other technologies<br>– A support team is available when there is difficulty in managing the EDMS<br>– The organization provides adequate infrastructure<br>– Training lessons are provided |

(continued)

**Table 1** (continued)

| Work | Objectives | Origin of Factors | Outcomes |
|---|---|---|---|
| [25] | To design an authoritative survey tool based on stringent instrument development protocols | A study on factors influencing the adoption of electronic and record management system (EDRMS) | – The organization provides adequate infrastructure<br>– Training lessons are provided<br>– The EDMS improves work performance<br>– EDMSs are effective in achieving organizational objectives<br>– EDMSs are easy to learn and easy to control<br>– The EDMS avails a user-friendly system interface<br>– Colleagues recommend the use of the EDMS<br>– Subordinates support the use of the EDMS<br>– The EDMS can be integrated into other technologies |
| [13, 23] | To explore factors contributing to EDMS implementation, with the objective of forming critical factors | A qualitative study of the Namibian public service | – Training for the employees<br>– Safety, security, and confidentiality of the records in the systems<br>– Risks of obsolescence of hardware and software<br>– System maintenance<br>– Workflow systems integration<br>– Top management support and resources commitment<br>– End use buy-in<br>– Processes of converting paper records to electronic format<br>– Ease of retrieving information via electronic sources compared to via paper records<br>– Ease of sharing information |
| [2, 15] | To examine factors that affect the use and non-use of the EDMS | A case study of the Ministry of Trade and Industry in Botswana | – EDMS functionality<br>– Usability and understanding of output<br>– Demonstration of benefits |

**Table 2** Critical success factors and definition of the UTAUT

| UTAUT factors | Critical success factor | Works |
|---|---|---|
| Performance expectancy | 1. EDMS functionality<br>2. Usability and understanding of output<br>3. Demonstration of benefits<br>4. Piloting and testing<br>5. Integration of systems and technology<br>6. User friendliness of EDMS<br>7. Efficiency of EDMS<br>8. Effectiveness of EDMS | [2, 11–14, 22, 24–26] |
| Effort expectancy | 1. Process readiness<br>2. Infrastructure readiness<br>3. Architecture readiness<br>4. Support of team-based approaches to problem solving<br>5. Spirit of cooperation and teamwork<br>6. Sharing of expertise<br>7. Change of management<br>8. Assurance that a project has a clear agenda<br>9. Alignment of projects with business objectives<br>10. Communication<br>11. Policies and guidelines | [6, 7, 12, 24, 25] |
| Social influence | 1. Management, leadership, and commitment toward EDMS<br>2. Colleagues' recommendation of the use of EDMS<br>3. Top management's recommendation of the use of EDMS<br>4. Subordinates' support of the use of EDMS<br>5. Impact of EDMS on reputation<br>6. High regard of people using EDMS<br>7. Planning and project management<br>8. Gaining of commitment and support of chief executive officers<br>9. Top management encouragement toward informal or formal communication<br>10. Development of clear mission regarding business objectives<br>11. Top management encouragement toward use of EDMS | [3, 10, 24, 25] |

(continued)

**Table 2** (continued)

| UTAUT factors | Critical success factor | Works |
|---|---|---|
| Facilitating conditions | 1. Active encouragement of employee participation in EDMS-related decisions<br>2. Involvement of all levels within the organization and external stakeholders<br>3. Consistent updates of management knowledge<br>4. Involvement of EDMS end users<br>5. Adequate training and support for users<br>6. Requirement-driven procurement planning<br>7. Provision of technical resources (e.g., equipment, software)<br>8. Human resource availability<br>9. Sufficient monetary resources provided to support EDMS implementation<br>10. Prior development or existence of necessary infrastructures | [7, 13, 23–27] |

## 4 Conclusion

The factors affecting government employees' acceptance of EDMSs were identified based on the literature review, and their relative importance was determined by prioritizing them using the AHP. The AHP is a multi-criteria decision-making (MCDM) tool that measures the importance of these factors through pairwise comparisons [21]. Based on the results obtained, 40 factors have been identified and categorized under four main variables of the UTAUT framework, namely, "Performance Expectancy (PE)," "Effort Expectancy (EE)," "Social Influence (SI)," and "Facilitating Conditions (FC)." The findings of this study reveal that "Performance Expectancy (PE)" and "Effort Expectancy (EE)" are the two most important factors that affect the use of EDMS among government employees.

The critical success factor from the literature is a means to align information technology planning with the strategic direction of an organization. However, this research has limitations that require further investigation. One limitation is that the rating scale used in the AHP is conceptual with a risk of bias while making pairwise comparisons of different factors [29]. Future work should take due care when deciding on the relative scores of the different factors. Also, the AHP technique may have excluded certain interrelationships among factors and sub-factors that need to be considered. Therefore, this study can be further extended by considering the adoption of other MCDM tools such as the analytic network process (ANP) [29, 30].

Besides that, most of the organizations may exploit the information that they have already entered into their electronic document management systems (EDMS) to extract hidden pattern or knowledge which is known as "Data Mining". Extracting hidden knowledge and summarizing data extracted from the EDMS is another area that should be exploited in order to review the acceptance of most organizations in using machine intelligence of the Fourth Industrial Revolution (4IR) technologies in managing their daily data manipulation operations for structured [31–33] and unstructured data [34–36].

## References

1. Radzi MA, Yatin SF, Fadzil NA, Aziz SA (2018) Document preparation for electronic document management system. Int J Acad Res Bus Soc Sci 8(9):179–190
2. Mosweu O, Bwalya KJ, Mutshewa A (2016) A probe into the factors for adoption and usage of electronic document and records management systems in the botswana context. Inf Dev 33(1):97–110. https://doi.org/10.1177/0266666916640593
3. Balogun NA, Raheem LA, Abdulrahaman MD, Balogun UO (2019) Adoptability of electronic document management system in Ilorin businesses. Niger J Technol 38(3):707. https://doi.org/10.4314/njt.v38i3.24
4. Rosa AT, Pustokhina IV, Lydia EL, Shankar K, Huda M (2019) Concept of electronic document management system (EDMS) as an efficient tool for storing document. J Criti Rev 6(5):85–90
5. McLeod J, Childs S, Hardiman R (2011) Accelerating positive change in electronic records management: headline findings from a major research project. Arch Manuscr 39(2):66–94

6. Kassab MI, Naser SSA, Al Shobaki MJ (2019) The role of policies and procedures for the electronic document management system in the success of the electronic document management system in the palestinian pension agency. Int J Acad Mult Res (IJAMR) 3(1):43–57
7. Abu Naser SS, Kassab MI, Al Shobaki MJ (2017) The impact of senior management support in the success of the e-DMS. Int J Eng Inf Syst (IJEAIS) 1(4):47–63
8. Kassab MI, Abu Naser SS, Al Shobaki MJ (2017) The impact of the availability of technological infrastructure on the success of the electronic document management system of the palestinian oension authority. Int J Eng Inf Syst (IJEAIS) 1(5):93–109
9. Al Shobaki MJ, Abu Amuna YM, Abu Naser SS (2016) The impact of top management support for strategic planning on crisis management: case study on UNRWA—gaza strip. Int J Acad Res Dev 1(10):20–25
10. Ambira CM, Kemoni HN, Ngulube P (2019) A framework for electronic records management in support of e-government in Kenya. Rec Manag J. https://doi.org/10.1108/rmj-03-2018-0006
11. Mukred M, Yusof ZM, Alotaibi FM, Mokhtar UA, Fauzi F (2019) The key factors in adopting an electronic records management system (ERMS) in the educational sector: a UTAUT-based framework. IEEE Access 7:35963–35980
12. Abidin SSZ, Husin MH (2018) Improving accessibility and security on document management system: a Malaysian case study. Appl Comput Inform. https://doi.org/10.1016/j.aci.2018.04.002
13. Karlos AN, Nengomasha CT. (2018). Change management: a critical factor for successful implementation of an electronic document and records management system (EDRMS): A namibian case study. University of Namibia. Retrieved from https://repository.unam.edu.na/handle/11070/2421
14. Su T, Wang SX, Chen Y, Tsou T, Cheng J (2017) Investigating the usability of electronic document management systems in government organizations from a human factor engineering perspective. J Adv Manag Sci 5:14–17
15. Mosweu O, Bwalya K, Mutshewa A (2016b) Examining factors affecting the adoption and usage of document workflow management system (DWMS) using the UTAUT model. Rec Manag J 26(1):38–67
16. Kruchinin SV, Bagrova EV (2019). Systems of electronic document management in Russian education: Pros and cons. In: 2019 international conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS), pp 628–630
17. Howard R, Restrepo L, Chang C-Y (2017) Addressing individual perceptions: an application of the unified theory of acceptance and use of technology to building information modeling. Int J Proj Manage 35:107–120
18. Madigan R, Louw T, Dziennus M, Graindorge T, Ortega E, Graindorge M, Merat N (2016) Acceptance of automated road transport systems (ARTS): an adaptation of the UTAUT model. Transp Res Procedia 14:2217–2226
19. Bozorgkhou N (2015) An Internet shopping user adoption model using an integrated TTF and UTAUT: evidence from Iranian consumers. Manage Sci Lett 5:199–204
20. Afshan S, Sharif A (2016) Acceptance of mobile banking framework in Pakistan. Telematics Inform 33:370–387
21. Saaty TL (1980) The analytic hierarchy process. McGraw-Hill, New York
22. Artamonov A, Ionkina K, Tretyakov E, Timofeev A (2018) Electronic document processing operating map development for the implementation of the data management system in a scientific organization. Procedia Comput Sci 145:248–253. https://doi.org/10.1016/j.procs.2018.11.053
23. Nengomasha C, Chikomba A (2018) Status of EDRMS implementation in the public sector in Namibia and Zimbabwe. Rec Manag J 28:252–264
24. Aziz AA, Yusof ZM, Mokhtar UA (2019). Electronic document and records management system (EDRMS) adoption in public sector instrument's content validation using content validation ratio (CVR). J Phys Conf Ser 1196, 1–7. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1196/1/012057/pdf

25. Aziz AA, Yusof ZM, Mokhtar UA, Jambari DI (2019) The intention to adopt electronic document and records management system: questionnaire development procedure. Int Conf Electr Eng Inf (ICEEI) 2019:590–595
26. Aziz AA, Yusof ZM, Mokhtar UA, Jambari DI (2018) A conceptual model for electronic document and records management system adoption in Malaysian public sector. Int J Adv Sci Eng Inf Technol 8:1191–1197
27. Wu CH, Chiu RK, Yeh HM, Wang DW (2017) Implementation of a cloud-based electronic medical record exchange system in compliance with the integrating healthcare enterprise's cross-enterprise document sharing integration profile. Int J Med Inf 107:30–39. https://doi.org/10.1016/j.ijmedinf.2017.09.001
28. Saaty TL (1990). How to make a decision: the analytic hierarchy process. Eur J Oper Res 48, 9e26. https://doi.org/10.1016/0377-2217(90)90057-I
29. Gupta KP, Bhaskar P, Singh S (2017). Prioritization of factors influencing employee adoption of e-government using the analytic hierarchy process. J Syst Inf Technol
30. Gokhale M (2007). Use of analytical hierarchy process in university strategy planning. Master's Theses. 4608. Retrieved from https://scholarsmine.mst.edu/masters_theses/4608
31. Alfred R (2008) DARA: data summarisation with feature construction. In: 2008 second Asia international conference on modelling & simulation (AMS), Kuala Lumpur, 2008, pp. 830–835. https://doi.org/10.1109/AMS.2008.131
32. Alfred R (2010) Feature transformation: a genetic-based feature construction method for data summarization. Comput Intell 26:337–357. https://doi.org/10.1111/j.1467-8640.2010.00362.x
33. Alfred R. (2007) The study of dynamic aggregation of relational attributes on relational data mining. In: Alhajj R, Gao H, Li J, Li X, Zaïane OR (eds) Advanced data mining and applications. ADMA 2007. Lecture notes in computer science, vol 4632. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73871-8_21
34. Alfred R, Mujat A, Obit J.H. (2013) A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles. In: Selamat A, Nguyen NT, Haron H (eds) Intelligent information and database systems. ACIIDS 2013. Lecture notes in computer science, vol 7803. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36543-0_6
35. Basri SB, Alfred R, On CK (2012) Automatic spell checker for Malay blog. In: 2012 IEEE international conference on control system, computing and engineering, Penang, 2012, pp. 506–510. https://doi.org/10.1109/ICCSCE.2012.6487198
36. Alfred R et al (2013) A rule-based named-entity recognition for malay articles. In: Motoda H, Wu Z, Cao L, Zaiane O, Yao M, Wang W (eds) Advanced data mining and applications. ADMA 2013. Lecture notes in computer science, vol 8346. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-53914-5_25

# Hadith Arabic Text Classification Using Convolutional Neural Network and Support Vector Machine


**Irwan Mazlin, Izani Mohamed Rawi, and Zaki Zakaria**

**Abstract** There are a lot of work has been implemented to solve the problem of text classification but There is only few researchers doing Arabic text classification because of the difficulties in text preprocessing. Convolution Neural network and support vector machine is two different algorithm that can be applied on text classification. CNN seems to be good in extracting the feature from input and SVM is good for classify the class. This study is to introduce Hadith text classification using Convolutional Neural Network and Support Vector Machine. In order to get preliminary result, we used BBC news article (English language) and Arabic tweet sentiment (Arabic language) as dataset for CNN with SVM model. There are 4 methods to evaluate the model which are f1-score, precision and recall and accuracy and error rate probability. We evaluate the model using accuracy and loss using different learning rate. The model accuracy and loss for preliminary result of BBC news article (English language) and Arabic tweet sentiment(Arabic language) are 0.857 accuracy, 0.245 loss and 0.884 accuracy, 0.344 loss. This shows that the proposed model has potential for Hadith text classification.

**Keywords** Artificial intelligent · Machine learning · Deep learning · Text classification · Hadith

## 1 Introduction

In this era of technology, text classification has become hot topic to most researchers due to the increasing data on the internet [1]. Most of researchers take the initiative

I. Mazlin (✉) · I. M. Rawi · Z. Zakaria
Faculty of Mathematical and Computer Science, Universiti Teknologi Mara, 40450 Selangor, Malaysia
e-mail: irwanmazlin@gmail.com

I. M. Rawi
e-mail: izani@tmsk.uitm.edu.my

Z. Zakaria
e-mail: zaki@tmsk.uitm.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_30

to do some work on text classification in order to get high accuracy and can solve text classification problem [2]. Text classification is the assigning of text document or corpus into several classes [3]. The increasing data on the data is also one of the factors that the researcher like to do some work on text classification. Unlike English, there are only few researchers doing text classification for Arabic language [3]. This study is about to introduce the combining between two algorithm which are Convolutional Neural Network and Support Vector Machine. This studies also we improve the model by applying bath normalization in order to combat overfitting The idea of this study is the model extract the text feature using CNN and training the data based on the extracted feature using Multi-layer Perceptron and Support Vector Machines We are using BBC news articles (English language) and Arabic tweet sentiment (Arabic language) as dataset in order to get preliminary result before we proceed to use the propose model for Arabic language. The purpose of this study is to investigate has the potential to apply on Hadith Arabic text classification. We have tested several models such are CNN + SVM using BBC news articles (English language) and Arabic tweet sentiment (Arabic language) for classification. In the next section, we organize the explanation of this study as follow: in Sect. 2 we present the related work of CNN and SVM; in Sect. 3 we explain the architecture design of CNN and SVM; in Sect. 4 is the preliminary result we get after train the model and Sect. 5 is the conclusion to close this studies.

## 2 Literature Review

The purpose of Arabic text classification is to classify a document with text into several classes such as zakat and hajj [4]. Due to the increase text data in the internet, text classification has become a hot topic in their research. There are two methods to classify text into several classes. First is manually created a set of rules and it will become an expert system to classify the text. The result from the expert system shows higher accuracy. However, it consumes a lot of time for text classification process. Secondly, machine learning algorithm. By using machine learning algorithm will give best performance to classify text. There are many machine learning techniques that can be used for text classification. Most of the studies have solved the problem of text classification using dataset English language. However, there are few researcher breakthroughs the idea by applying text classification in Arabic language. For example, [5] proposed a project for text classification using convolutional neural network by using English language as the dataset; [6] presented a paper by using LSTM-CNN on text classification able to improve accuracy of the model; [7] did an experiment on text classification using convolutional neural network. They stated that by using this model, it has the ability to improve the accuracy. In [8], it is effective using convolutional neural network model in NLP and it also show higher accuracy in semantic classification. These all work use English language as dataset to train the model. However, there are few researchers applied text classification using support vector machine in Arabic language. For example, [9] propose using SVM

on Arabic text classification with active learning method; [10] proposed text classification using SVM and they did the experiment by segmented the word on Arabic text classification; In [11] did an experiment by compare SVM technique with other technique in Quranic text classification; [12]; [13] proposed a project using SVM on Arabic text classification and employ N-gram kernel in SVM.

## 3 Methodology

This study proposed a model can classify single-label data. This propose model use Kitab bukhari, muslim and tirmidhi dataset in Arabic language. The dataset has 1231 documents with 3 different classes which are prayer, zakat and fasting.

In the course of this study, the model should be able to give high accuracy in the end of model training and classify the model has several main processes which are data collection, text preprocessing, data separating into training and testing, and classification process. In general, Fig. 1 illustrates the process of this study.

- Data Collection and Preparation

In this research, the researcher using CNN + SVM to classify the topic of Hadith Bukhari in Arabic text classification. The researcher will collect and classify manually the selected topic in Hadith which are Prayers, Fasting and zakat. The data can be collected at different sources such as sunnah.com and http://islamport.com/Al-Bokhary.pdf. However, due to the limited of time, the researchers use BBC news article (English language) and Arabic tweet sentiment (Arabic language) investigate whether the model has the potential for Hadith Arabic text classification. The BBC news article (English language) and Arabic tweet sentiment (Arabic language) consist four classes which are sports, politics, economic, and entertainment. The total text in the document is 17 k text.

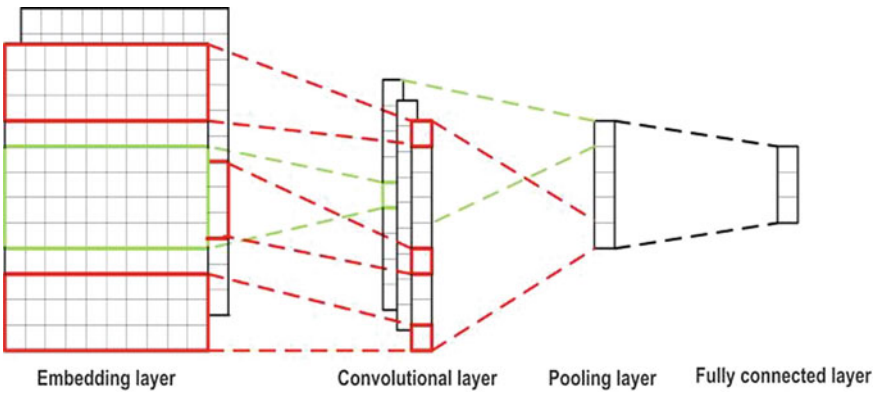



**Fig. 1** Convolutional neural network architecture for text classification [7]

- Text Pre-processing

Text Pre-processing can be done by removing noise in contained dataset. In this study, the researcher uses several, text pre-processing method namely Tokenization, noise cleaning, stopword removal and stemming. The complete pre-processing step shown in Fig. 2.

- Tokenization is the process of convert word into unique number. This process uses the existing library function in python programming language.
- Cleaning is the processing of removing punctual mark, symbol, numbers, and non-arabic character that exist in the data. This process uses the existing library function in python programming language
- Stopword is the process of removing the word that are not related information in the dataset.
- Stemming is the process removing prefix and suffix in word to convert it into base word. This process uses the existing library function in python programming language
- Normalization is the process of normalize the character in word. For example, the huruf of taa marbutah in alif, and different form of alif ( ٱ , إ , أ) convert into ا.This process uses python programming language to normalize the Arabic character

After finish cleaning the data from noise, the data will split it into training sample and testing. This study, the researcher split it into 80% for training sample and 20% for testing sample.

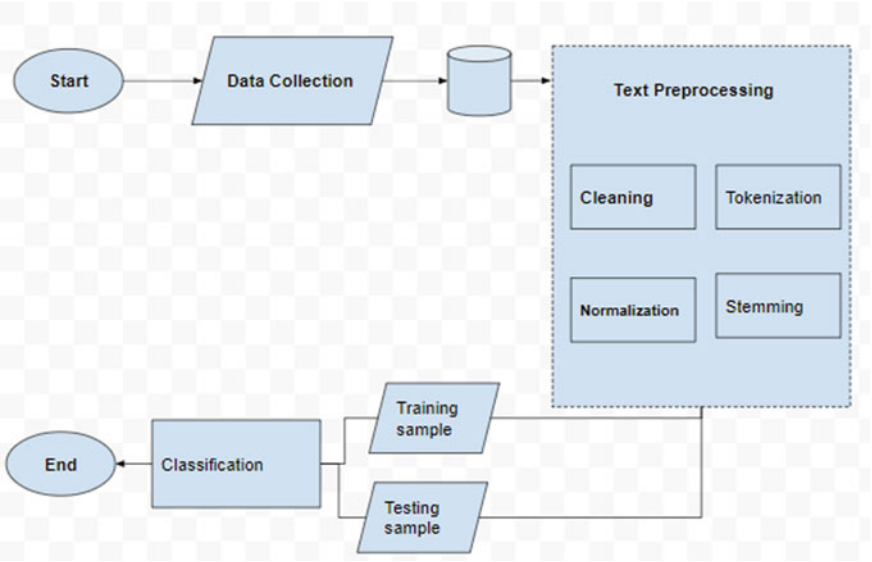



**Fig. 2** Flowchart of Hadith Arabic text classification

- Classification using CNN and SVM

In this research, we use CNN for classification because the complexity of Arabic language make CNN is a good candidate for classification [1]. The researcher also implement the Batch Normalization in CNN to counter overfitting. However, the researcher eliminates the dropout in CNN model since BN also can be act as one of regularization technique to counter overfitting. CNN model for Hadith Arabic Text Classification consist of four layer with are Embedding layer, Convolutional Layer, Rectified Linear Unit, Pooling Layer, and Full connected layer. The researcher using SVM as classifier in this researcher. The Fig. 1 shows the architecture of CNN.

## *3.1 Embedding Layer*

Embedding layer is the text data is vectorize to form matrix. Each row is a vector and represent a word. Each row of the matrix also represent as a token, can be a word or character. The researcher using a word to represent a token in embedding layer. The representation of text data can be made by using equation below.

$$s \times d \tag{1}$$

## *3.2 Convolutional Layer*

Convolutional layer is the layer that will do the convolution operation over the text matrix. Convolution operation is the multiplication and addition between the filter and text matrix. The larger the number of filter use for convolution operation produce a lot of feature maps in convolutional layer. The width of the filter is equal as the length of the sentences. By convolving the filter over the width and height of the input, feature maps can be generated [14, 15]. The equation below is the formula for convolution process:

$$\sum_{x=1}^{k} g(y \cdot d - x + c) \tag{2}$$

### 3.3 Rectified Linear Unit

Rectified Linear Unit is the activations function in Convolutional Neural Network Model. The goal of rectified linear unit is to replace negative value with 0 within input matrix and any positive value in the input matrix will remain the same. The operation of Rectified Linear Unit can be applied by using equation and the graph of rectified linear unit is shown below:

$$f(x) = \max(0, \text{input}) \tag{3}$$

### 3.4 Pooling Layer

By apply pooling layer, it able to reduce the size of feature maps in rectified linear unit. There are two types of pooling layer which are compute max value of convolution or average value of operation [16]. If we used max pooling, we are only taking the maximum value in the input feature maps [17] and average pooling is to calculate the average value in the matrix of feature map [18]. The equation below is the equation for pooling process

$$\max(g(y \cdot d - x + c)) \tag{4}$$

### 3.5 Multi-layer Perceptron

Typically, fully connected layer was used at then of convolutional neural network. Fully connected layer deals with the combination of feature in order to predict classes. The process of fully connected layer is the value of the output in the last layer of convolutional neural network works as the input in fully connected layer and it will determine which feature are belong to particular class. The operation of fully connected layer is the layer will compute the product of weight and adjust weight to correct the probabilities for each class [19]. The way of the model adjusts the weight is through backpropagation process [20]. Basically, the purpose of backpropagation process is to find the most accurate weight for each class. Each neuron will have their own weight and their weight will be adjusted. Usually, the researchers use Softmax as an activation function in MLP to adjust weight to predict classes [21]. However, we are using Hinge loss function in order to use SVM as classifier to predict classes. The way the model classifies the classes is using voting process. Every neuron vote on every label and the most vote of weight is the decision of class. The operation of backpropagation can be made using formula below.

### 3.6 Batch Normalization

Batch normalization is one of the techniques to avoid overfitting. There are several ways to avoid overfitting which are dropout and regularization. In this study, the researcher used Batch Normalization to avoid overfitting. Batch Normalization is a technique to solve the problem of the model on overfitting in a model. The purpose of using batch normalization is to reduce covariate shift in the network [14, 15]. Covariate shift is the change of the distribution of the hidden activation. For example, during the training of neural network, if the distribution of hidden activation has a change in weight and bias in current layer, it will changes for the next layer too. In this paper [17], they are using batch normalization in the network and it helps the network to train efficiently. The Fig. 3 shows the scenario of model using batch normalization and without using batch normalization.

The benefit of using Batch Normalization is it able to accelerate model to learn data and it improve the accuracy of the model [14, 15]. It is also can reduce the dropout usage because by using batch normalization can speed up model training without overfitting. The equation below shows the batch normalization algorithm:

$$\mu\beta \leftarrow \frac{1}{m}\sum_{t=1}^{m} x_1 \tag{5}$$

$$\sigma_\beta^2 \leftarrow \frac{1}{m}\sum_{t=1}^{m} (x_1 - \mu\beta)^2 \tag{6}$$

$$\hat{x} \leftarrow \frac{x_1 - \mu\beta}{\sqrt{\sigma_\beta^2 + \in}} \tag{7}$$

$$y_1 \leftarrow \gamma\widehat{x_1} + \beta \equiv BN_{\gamma_1\beta}(x_1) \tag{8}$$

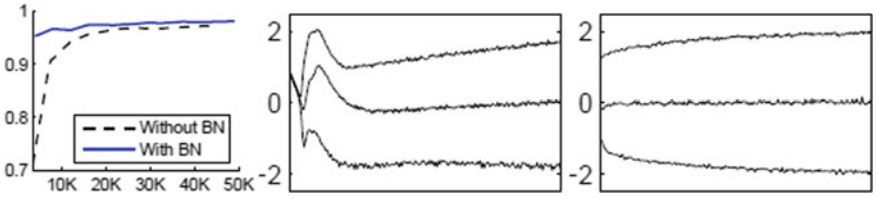



**Fig. 3** Scenario with and without using batch normalization

### 3.7 Support Vector Machine

Support vector machine is a machine learning algorithm where the model analyses the data for classification and regression task. SVM is a machine learning algorithm that introduced by [18]. Basically, SVM constructs hyperplane in high dimensional space for classification and regression task. If the hyperplane has large distance of training data point, it should achieve good separable between classes [22]. The main goal of SVM is to find the optimal separating hyperplane that has maximum margin between two classes. The larger the margin, the lower the error rate of the model [23]. SVM classifier can be applied in several areas which are text classification, image classification and regression. From the theory of computational learning theory, SVM construct the hyperplane to separate into two classes and SVM construct the hyperplane based on support vector. The SVM process can made using equation below:

$$\min \frac{1}{p} w^T w + C \sum_{i=1}^{p} \max\left(0, 1 - y_i'\left(w^T x_i + b\right)\right) \tag{9}$$

Figure 4 shows the process of Hadith Arabic Text Classification for this study. This study is to proposed a model for Hadith Arabic text classification using Convolutional Neural Network and Support Vector Machine. However, due to the limited of time, the researcher used BBC news articles and Arabic tweets sentiment to investigate the potential of the model for this research. First step is embedding layer to represent text into input matrix. The text data in vectorized to form in matrix. Each row of vector represents a token which is a word. The dimensional word is 300. Then, the matrix will become an input in CNN to generate feature maps. The value of weight in pooling layer will become an input in full connected layer. For this pre-test model. This research using multiple size of filter to generate feature maps with (4, 5, 6) since in this research [7] using this parameter and achieve higher accuracy for

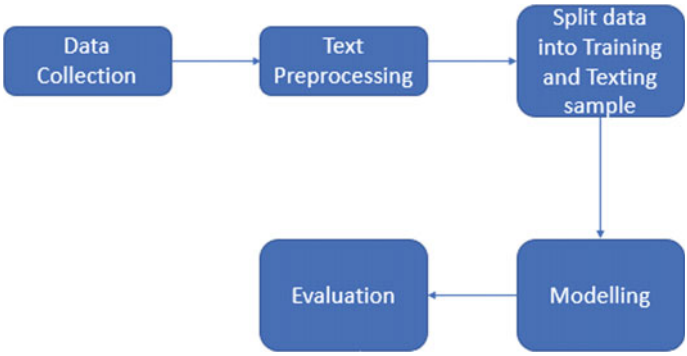


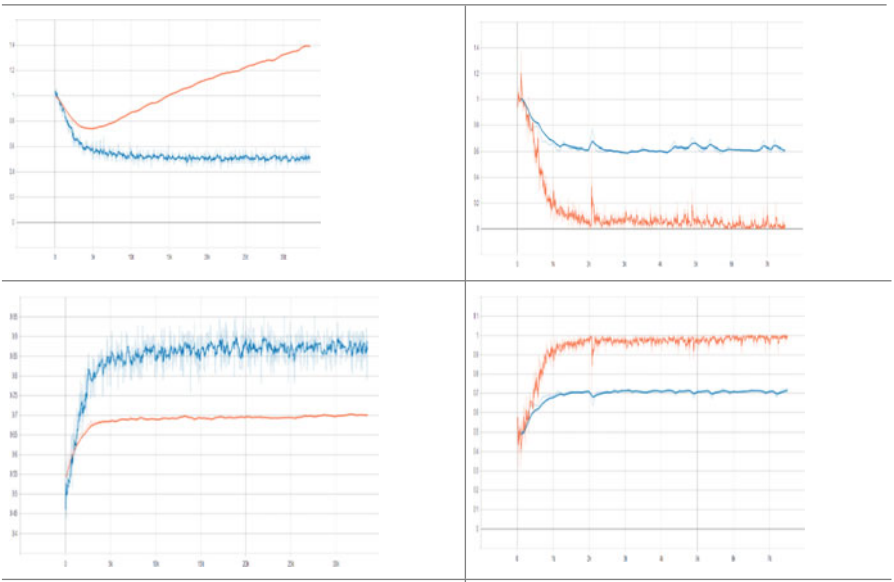

**Fig. 4** The process of Hadith Arabic text classification

their research. However, the researcher did not used dropout to counter overfitting because based on this study [14, 15], BN also can be act as regularization technique to counter overfitting. The total number of filters is 50. Finally, for fully connected layer, the researcher applied 32 units in order to perform 1 of 2 classes for Arabic tweet sentiment and 1 of 5 classes for BBC news articles followed by the SVM classifier.

## 4 Result and Discussion

We have tested different learning rate value and apply batch normalization and without batch normalization in the network. Different parameter testing gives different result of accuracy. We used four layers of convolutional and pooling and two layers of dense and SVM as classifier which is used hinge loss as calculation for error rate value. In order to train the model, we used tensorflow framework as a tool to develop the model. The framework able to set how many layers and set the parameter the fit for training. The researcher test the model with the parameter shown in Table 4. The parameter shown in Table 4 achieves up to 85.7% accuracy, 24.5% loss for English language, and 88.4% accuracy, 34.4% loss for Arabic language and regularization which is BN were applied in CNN after rectified layer. Based on our finding in Table 1, the model shows that it has the potential for Hadith Arabic text

**Table 1** Performance of the model



Loss and accuracy without BN | Loss and accuracy with BN

**Table 2** The evaluation model using different dataset

| Dataset | Accuracy | Loss |
|---|---|---|
| BBC news article(English Language | 85.7% | 24.5% |
| Arabic tweet sentiment (Arabic language) | 88.4% | 33.4% |

classification. Plus, the complexity of Arabic language make CNN is a good candidate to apply on Arabic language [4]. The Table 1 shows the performance of the model in view of accuracy and loss.

Table 2 below show the evaluation of the model using different dataset. The dataset used are BBC news article (English language) and Arabic tweet sentiment (Arabic language). The evaluation shows that the model gives reasonable result using both dataset in English language and Arabic language. The loss of the model without BN seems overfitting in middle of training phases and with BN continuous decreasing. It prove that BN also can be used as regularization techniques as mention in this research [5]. The contribution of this research is we eliminate the dropout in the architecture of CNN that usually people used combat overfitting like in this study [6, 7] and attach BN in the CNN layer to avoid overfitting (Tables 3 and 4).

**Table 3** Show the accuracy and loss graph using different dataset

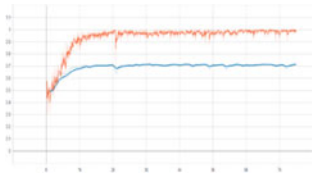
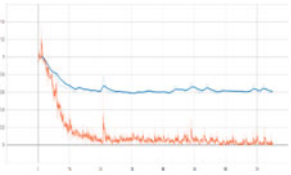
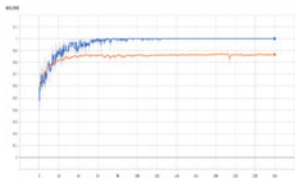
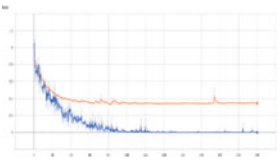
| | Accuracy | Loss |
|---|---|---|
| BBC news article(English Language) Blue line is testing and red line is training |  |  |
| Arabic tweet sentiment (Arabic language) Blue line is training and red line is testing |  |  |

**Table 4** Parameter setting

| | |
|---|---|
| Size of filter | (4, 5, 6) |
| Learning rate | 0.01 |
| Filter striding size | 2 |
| Dimension | 300 |
| Length of sequence | Maximum value of text |

## 5 Conclusion

In this research, we are using three different model on BBC news article (English language) and Arabic tweet sentiment (Arabic language) dataset to investigate the performance of this model for Hadith Arabic text Classification for future work. The result show that CNN + SVM model give reasonable accuracy with 85.7% for English language and 88.4% for Arabic language. It should give the promising result if apply on Hadith dataset. Besides that, we also will do comparison with the existing word representation model for Hadith Arabic text classification.

**Acknowledgements** This work is supported by Faculty of Mathematical and Computer Science. The authors would like also to thank Universiti Teknologi Mara for its support of this research work.

## References

1. AKBEN (2019) Arabic text classification using polynomial networks. J King Saud Univ-Comp Inform Sci 23(3):1–19
2. Onan A, Korukoğlu S, Bulut H (2016) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57:232–247
3. Goudjil M, Koudil M, Hammami N, Bedda N, Alruily M (2013) Arabic text categorization using SVM active learning technique: an overview. In: 2013 World Congress on Computer and Information Technology WCCIT 2013, pp. 7–8
4. Shahrul M, Sharifuddin I, Nordin S, Ali AM (2020) Comparison of CNNs and SVM for voice control wheelchair. IAES Int J Artif Int 9(3):387–393
5. Banerjee I et al. (2018) Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med, November
6. Wang X, Li J, Liu Y (2019) Application of convolutional neural network (CNN) in microblog text classification. In: 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing ICCWAMTIP, pp. 127–130
7. Zhao F, Gaw SD, Bender N, Levy D (2016) Exploring cloud computing adoptions in public sectors: a case study. GSTF J Comput 3(1):1–10
8. Kim Y(2011) Convolutional neural network for sentences classification
9. Yih W, He X, Meek C (2014) Semantic parsing for single-relation question answering. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, vol. 2: Short pap., no. May, pp. 643–648
10. Yih W-T, Toutanova K, Platt JC, Meek C (2011) Learning discriminative projections for text similarity measures. In: Proceedings of the fifteenth conference on computational natural language learning, June, pp. 247–256
11. Al-Thubaity A, Al-Subaie A (2016) Effect of word segmentation on Arabic text classification. In: Proceeding 2015 International Conference on Asian Language Processing IALP, pp. 127–131
12. Al-Kabi MN, Ata BMA, Wahsheh HA, Alsmadi IM (2013) A topical classification of Quranic Arabic text. In: Taibah university international conference on advances in information technology for the holy Quran and its sciences, December, pp. 272–277
13. Al-anzi FS (2019) Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Comput Inf Sci, pp. 1–19

14. Abdel-hamid O, Jiang H, Penn G (2012) Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. Department of Computer Science and Engineering, York University, Toronto, Canada, Department of Computer Science, University of Toronto, Toronto, Canada, pp. 4277–4280
15. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift
16. Zin IA, Ibrahim Z, Isa D, Aliman S, Sabri N (2020) Herbal plant recognition using deep convolutional neural network. Bull Electr Eng Inf 9(5):2198–2205
17. Windows M, Corporation M, Hori K, Sakajiri A efficient back-prob
18. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 297:273–297
19. Zaini N, Malek MA, Yusoff M, Fatimah S, Osmi C, Mardi NH (2018) Support vector machine and neural network based model for monthly stream flow forecasting, vol. 7, pp. 683–688
20. Nehar A, Ziadi D, Cherroun H (2016) Rational kernels for Arabic root extraction and text classification. J King Saud Univ—Comput Inf Sci 28(2):157–169
21. Chen G, Parada C, Heigold G (2014) Small-footprint keyword spotting using deep neural networks. In: IEEE International conference on acoustics, speech and signal processing, i, pp. 1–5
22. Fatima S (2017) Text document categorization using support vector machine. Int Res J Eng Technol 4(2):141–147
23. Liu Z, Lv X, Liu K, Shi S (2010) Study on SVM compared with the other text classification methods. In: 2nd International Workshop on Education Technology and Computer Science ETCS 2010, vol. 1, pp. 219–222

# Alice: A General-Purpose Virtual Assistant Framework


**Soon-Chang Poh, Yi-Fei Tan, Chee-Pun Ooi, Wooi-Haw Tan, Albert Quek, Chee-Yong Gan, Yew-Chun Lee, Zhun-Hau Yap, and Chin-Leei Cham**

**Abstract** In this paper, a virtual assistant framework called Alice is presented. This virtual assistant is a combination of 3D avatar, face detection, face recognition and face expression recognition with a voice assistant that similar to Amazon's Alexa. The 3D avatar (Alice) is a female character animated using Unity and the lip is animated to sync with the speech to make it looks like speaking. Besides that, the 3D avatar can display different facial expressions such as happy, sad and upset. Face detection and recognition makes the system aware of the human user's identity. Whereas, face expression recognition enables the system to detect the facial expression of the human user. Whenever there is a question being asked, the system will use Speech-to-Text system to convert human speech to text and Natural Language Processing to interpret the intent behind the text. Based on the result of interpretation, the system decides which audio file to be used as response. Then, a realistic artificial voice is generated as response to the human user. The system can access database based on user's identity to retrieve information about that user. This may create a personalized experience for the human user. This framework can be customized for other applications for different fields. For this Alice framework, two applications have been developed namely a question answering chatbot and a customer service agent.

**Keywords** Virtual assistant · Face recognition · Face expression recognition · 3D avatar · Natural language processing and Speech-to-Text

S.-C. Poh (✉) · Y.-F. Tan · C.-P. Ooi · W.-H. Tan · A. Quek · C.-Y. Gan · Y.-C. Lee · Z.-H. Yap · C.-L. Cham
Multimedia University, Cyberjaya, Selangor, Malaysia
e-mail: psoonchang@gmail.com

Y.-F. Tan
e-mail: yftan@mmu.edu.my

C.-P. Ooi
e-mail: cpooi@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_31

## 1 Introduction

A chatbot is a software application which is used in place of a human staff to carry out conversation online through text or speech. Traditionally, online customer service and call centres staffed by human workers are responsible for this task. Due to the rise of Artificial Intelligence (AI), chatbots are becoming more intelligent and are taking over the task of human staffs [1, 2].

There are two types of chatbots which are text-based chatbots [3–5] and speech-based chatbots [6, 7]. Text-based chatbots interacts with users via text. They have many applications which include customer support and question answering. In [3], a conversational chatbot was developed to simulate a technical support representative to provide customer services for an online course provider. The developer's goal is to build a cost-effective and efficient chatbot to minimize the cost of customer service. Potential customers can use the chatbot via Twitter, Messenger, Telegram and WhatsApp. The message from the user of the chatbot is parsed using Natural Language Processing (NLP) [8] to understand user's intention in the input message. Based on the user intention, the chatbot retrieves pre-defined response or conducts search in historical data to construct the response. Chatbot is incapable of understanding every question and responds accordingly and this leads to poor experience for customer.

In [4], a text-based chatbot was developed to answer university-related questions. The message from the user is processed via several steps. Firstly, the program checks and fix the spelling of the words in the messages. Then, the program breaks down the message into words and check if the words exist in database. The authors defined several important keywords relevant to questions in the database. If the words matched, the program carries out a SQL query to retrieve the answer.

In [5], a chatbot was developed as an E-Learning assistant. The NLP component consists of two different NLP models which include a retrieval-based model and a QANet model [9]. The QANet model uses a neural network architecture which consists of convolutional neural networks (CNN) with self-attention. It is trained on Stanford Question Answering Dataset (SQuAD). The QANet is trained for reading comprehension question answering. When given a user's question and an article as input, the QANet will select parts of text from the given article as output. To use this model for chatbot, the authors modified the training method for QANet for everyday conservation. Besides that, psychologists and AI engineers have teamed up in building a chatbot for mental health therapy [5]. The chatbot is based on a type of widely used therapy called Cognitive Behavioral Therapy (CBT) [10]. A study by Stanford University shows that this chatbot significantly reduce anxiety and depression among people aged 18–28 years old when compared to an information-only control group [11].

On the other hand, speech-based chatbots often comes as an application on smart phone (Siri and Cortana) or standalone speaker devices such as Amazon's Alexa [6, 7]. These type of chatbots are also sometimes called virtual assistant or voice assistants. The main difference of these types of chatbots from text-based chatbots are that they can interpret human speech and provide response in form of synthesized

voice. Speech recognition is used for converting the speech of human user to texts. NLP is used to interpret the intention of human user in the texts. Then, the voice assistant will respond with synthesized voice. These voice assistants are used for various tasks such as home automation control, booking a flight ticket and managing to-do list.

In [12], the authors developed a chatbot with 3D avatar and text emotion recognition. The 3D avatar is a graphical representation of the virtual character which is built using Unity. The 3D avatar can display various expressions such as happy, sad, surprise and angry. A Long Short-term Memory (LSTM) neural network [13] was trained using dataset of conversations to generate response in form of text when given an input text. The generated text is then synthesized to voice and saved as an audio file. The audio is then played in sync with the 3D avatar.

In [14], the authors developed a 3D avatar which replicates what the user speaks in synthesized voice. The 3D avatar is a graphical representation of the user. In this work, the user speaks to a microphone and have his or her voice recorded. Speech recognition is used to convert the user's speech to text using HTML 5 Speech Recognition API. Then, Text-to-Speech (TTS) system based on HTML 5 Speech Synthesis API converts the text to speech. Then, the synthesized speech is played in sync with the 3D avatar to make it seems as if 3D avatar speaks. However, the author observed that speech recognition is not robust for distinguishing user's speech.

A survey concluded that rule-based type chatbots for marketing can provide simple and fast information [15]. However, customers have concerns that the chatbot might be inaccurate. The advantage of using chatbots in business is that they can siphon off frequent simple tasks from human staff. Compared to text-based chatbots, conversation with speech-based chatbots are more natural. On the other hand, speech-based chatbots raise privacy concerns. They are basically connected speakers which continuously record, and upload recorded audio to the server for data processing. Besides that, all these types of chatbot lack personal touch because they are not aware of the identity of the user.

In this paper, a voice assistant with 3D avatar for virtual character [12–14], Facial Expression Recognition (FER) [14] and face recognition was developed. The virtual assistant is aware of user's identity and facial expression using face recognition and FER respectively. The virtual assistant was developed as a general-purpose virtual assistant framework which can be customized for applications across different fields. A question inquiry assistant and customer service assistant application were developed using this proposed framework as proof of concept.

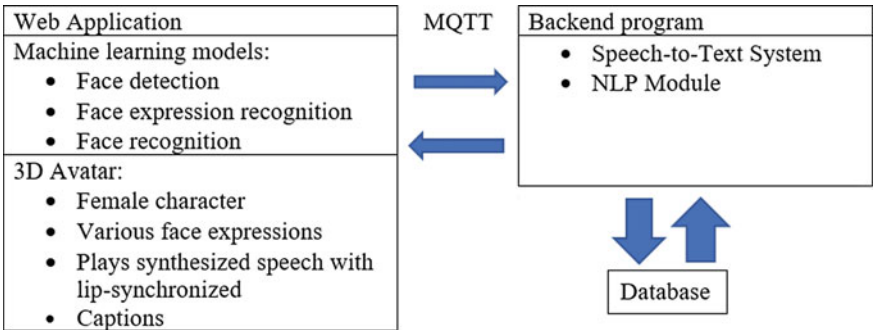



**Fig. 1** Overview of the system design

## 2 System Design

### *2.1 Overview*

As shown in Fig. 1, the Alice virtual assistant framework consists of two main parts namely web application and backend program. The web application consists of two main parts namely machine learning models and the 3D avatar. The machine learning models include face detection model, face expression recognition model and face recognition model. The 3D avatar is a female virtual character which can display various face expressions. Besides that, the web application plays the synthesized speech with lip-synchronized with the 3D avatar. In addition, caption for the speech is displayed. On the other hand. the backend program consists of two main parts namely Speech-to-Text (STT) System and NLP module. The backend program accesses and store information in the database based on the user's identity. The different parts communicate with each other via a protocol called Message Queuing Telemetry Transport (MQTT).

In Sect. 2.2, the 3D avatar of the web application is discussed in details. Section 2.3 presents the machine learning models used by the web application. Section 2.4 discusses the backend program's Speech-to-Text system and NLP module. Section 2.5 presents the process flow of the Alice framework.

### *2.2 3D Avatar*

The 3D avatar is a female virtual character called Alice. It is built using Unity, a cross platform game engine which can be used for 3D visualization. The 3D avatar is designed to have multiple facial expression such as happy, sad and upset as shown in Fig. 2.

**Fig. 2** 3D avatar's facial expressions: happy, sad, upset

The various expression is displayed to enhance user experience when interacting with Alice. Alice can response to human user using synthesized speech generated by Google's Text-to-Speech (TTS) system. Google's TTS system uses a generative model called WaveNet to generate speech that sounds more natural with more human-like characteristic on syllables, phonemes and words. The audio file can be played lip-synchronized with the 3D avatar. The lip-synchronizing is enabled using a tool called Rhubarb Lip Sync [16] which analyzes audio files and generate lip-sync information. It can be used to animate the lip of the 3D avatar.

Figure 3 shows the web application which consists of 3D avatar running on Google Chrome web browser. Beneath the virtual character, the caption for the speech that Alice is speaking is displayed.


**Fig. 3** Web application for Alice framework

### 2.3 Machine Learning Models

In this framework, three machine learning models for face detection, face recognition and face expression recognition are included. These models are implemented on the web application using *face-api.js* library [17]. Face detection model outputs the location of faces in a given image. The location of the face is given in the form of bounding box with four parameters as listed in Table 1. There are two face detection models available in the *face-api.js* library namely SSD MobileNet V1 and Tiny Face Detector. The Tiny Face Detector was selected because it is smaller and has similar accuracy. It was trained on a training set which consists of around 14,000 images annotated with bounding boxes.

The faces in the image are cropped and then passed to the face recognition model as input. The face recognition model learns to map faces to a 128-dimensional vector space. For each of the face images, face recognition model generates a 128-dimensional vector called face embedding vector. The distance of the face embedding vectors correspond to similarity in faces. The face recognition model is trained to generate face embedding vectors that are close to each other for face images of the same person and generates face embedding vectors which are far away from each other for face images of different person. The *face-api.js* library uses a face recognition model based on ResNet architecture. The face recognition model was trained by Davis King and achieved an accuracy of 99.38% on the Labeled Faces in the Wild (LFW) dataset [18]. Face recognition model architecture is some variant of deep convolutional neural networks [19, 20]. There are a few open source face recognition models that are publicly available such as FaceNet [21], VGGFace2 [22] and ResNet [18]. ResNet face recognition model was selected because it can be easily integrated with the web application using *face-api.js*.

A measure of similarity is used to gauge the similarity of two face embedding vectors. The measure of similarity used is Cosine Similarity. Cosine Similarity measures distance between two vectors using inner product. If there are two vectors denoted by A and B, then the Cosine Similarity can be calculated using formula (1). The value of Cosine similarity is in the range of [0, 1] and larger value means the vectors are more similar.

$$\text{Cosine Similarity} = \frac{\mathrm{A} \cdot \mathrm{B}}{\|\mathrm{A}\| \|\mathrm{B}\|} \tag{1}$$

**Table 1** Parameters of the bounding box

| Parameter | Description |
|---|---|
| $x$ | Top left $x$-coordinate of face |
| $y$ | Top left $y$-coordinate of face |
| $h$ | Height of the face |
| $w$ | Width of the face |

Deploying trained face recognition model consists of a few steps. Firstly, face images of people are collected and encoded into face embedding vectors. These face embedding vectors are arranged into a matrix and stored locally. Face detection model may detect many faces. However, the system performs face recognition on the largest face because it is the face of the person closest to the camera. Given a new face image, the face recognition model encodes it into a face embedding vector. The Cosine Similarity between the new face embedding vector with each of the stored face embedding vectors in the matrix are computed. Each of the stored face embedding vectors belong to a person. The identity of the given image will then be recognized as the owner based on Cosine Similarity value. If the value is more than 0.7, then new face will be recognized as one of the known faces. If it is less than 0.7, it is classified as an unknown person.

The face expression recognition model detects the expression when given a face image [23]. It can detect up to seven facial expressions which include surprised, disgusted, fearful, sad, angry, happy and neutral. According to *face-api.js* documentation, it was trained on publicly available dataset and images obtained from the web. The info about the largest bounding box, face expression and identity are published to a MQTT topic consecutively over time in JSON format.

## *2.4 Speech-to-Text System*

The backend program consists of two parts namely Speech-to-Text System and NLP module. Speech-to-Text (STT) system converts speech into text to be analyzed by the NLP module. For Alice framework, Google Cloud Speech-to-Text service is used. Google Cloud STT is one of the most accurate and reliable speech recognition services along with IBM Watson and Microsoft Azure services [24].

Figure 4 illustrates the process flow of the STT system of Alice framework. Google Cloud STT provide online speech recognition service. The audio is continuously stream to Google's server instead of sending it as an audio file after the recording is complete. The microphone's stream is closed if it is silence for more than 1 s. In this context, silence means the speech recognition model does not generate any output text for the audio because there is no human speech or the user speaks too softly. If Google Cloud returns text during the 1 s time period, the microphone will continue to stream until the third second. According to the Google Cloud user console's statistics, the median time taken for full text transcript to be sent back is 5.1 s. Then, the text transcript is passed to the NLP module. NLP module will interpret the intent behind the text and publish the response audio file to be played via MQTT. The entire process from microphone starts streaming to Alice responding take less than 6 s.

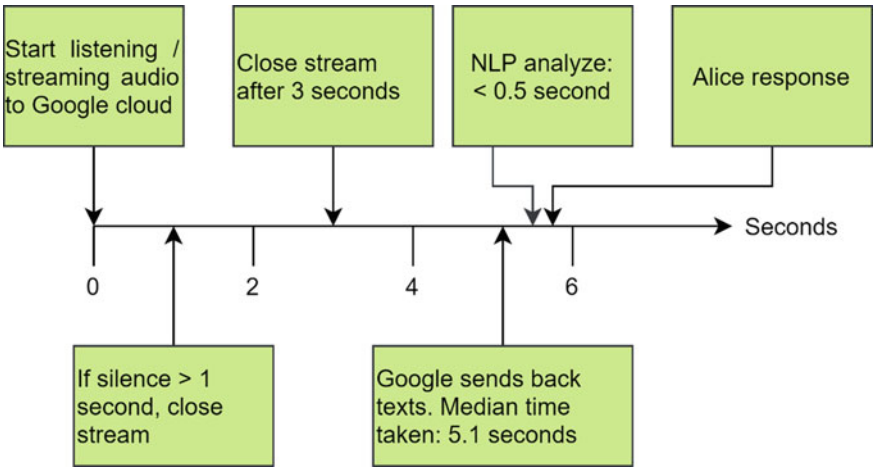



**Fig. 4** Design of Speech-to-Text system for Alice framework

## *2.5 NLP Module*

The Natural Language Processing (NLP) module is used to find the intent behind the text transcript. It consists of two layers namely deep learning model and keyword searching. The deep learning model used here is called Universal Sentence Encoder and the TensorFlow Hub library is used to implement Universal Sentence Encoder.

The working principle for Universal Sentence Encoder is similar to face recognition model. For face recognition model, the similarity of two images are calculated by using face embedding vectors. Whereas the Universal Sentence Encoder calculates the similarity of texts. The Universal Sentence Encoder will first encode the text into high-dimensional vectors. Then the inner product between 2 vectors is calculated to check the similarity of texts. The inner product of two encoded vectors is a scalar which fall in the range [0, 1]. If the value is larger, that means the text is more similar to the other text. The process of using Universal Sentence Encoder begin by preparing some reference sentences and encodes them into vectors. These vectors are then stored in computer memory. Secondly, whenever there is a new text transcript, the model encodes it into a vector. Inner product of the new vector with each of the reference vectors are calculated. If the largest inner product is more than 0.5, then the intent of the given new text is same as the reference sentence which is represented by that vector with the largest inner product. Then, backend program decides which audio file to be played as response. If the inner products are less than 0.5, the given text is classified as unknown. The unknown text is then passed to the second layer of the NLP module.

The second layer of the NLP module is keyword searching. This layer conducts search in the text transcript for pre-defined keywords using regular expression. If one of the keywords exists in the text, the intent behind the text can be determined. Then,

backend program can decide which audio file to be played as response. The backend program informs the web application which audio file to be played as response using MQTT.

## *2.6 Process Flow of Alice Framework*

Alice has two operating modes namely dormant mode and conversation mode. During dormant mode, most of the components of the backend program such STT system and NLP module are not active. This is because STT system is a usage-based billing service. Leaving it on all the time will incur a very high cost. During dormant mode, the face detection model is constantly detecting faces. After faces are detected on frame captured by the camera, the face image with the largest bounding box is selected and other face images are ignored. The largest face image is then passed to the face recognition and face expression recognition model. The combined process of face recognition and face expression recognition is tested in two different laptops. It takes less than 350 ms on a laptop with Intel i5 processor and a GTX1050 graphic card. The process takes less than 100 ms on a laptop with AMD Ryzen 5 processor and a RTX2080 graphic card.

After face recognition and face expression recognition are completed, the results are published to a MQTT topic. The backend program subscribes to that MQTT topic. Once there is a new update, Alice transitions into conversation mode. During conversation mode, the STT system starts. The identity of user will be recognized through face recognition model and the result from face expression recognition will determine the expression of the 3D avatar. The output of the face expression recognition is classified into two categories. Expressions such as surprised, disgusted, fearful, sad and angry are classified as negative expression. Whereas, happy and neutral are classified as positive expression. The 3D avatar will display sad expression when playing audio file if negative expression is detected and display happy expression when playing audio file if positive expression is detected. After NLP module has analyzed the intent behind the text, the audio file is determined. If the intent of the user requires access to database, a SQL query based on the intent and the human user's identity is executed to retrieve necessary info. The backend program publishes the audio file and expression to be played by Alice to a MQTT topic. The web application subscribes to the MQTT topic. Once a message consists of audio file and expression to be played is received, the audio file is played, and the lip animation starts. After that, Alice transitions back to dormant mode.

## 3 Applications and Conclusion

In this work, two applications of the Alice framework which include question inquiry assistant and customer service assistant were developed. The question inquiry assistant is similar to basic question and answering (Q&A) provided by voice assistant such as Amazon's Alexa (Demo: https://www.youtube.com/watch?v=tdzAi6-A4kw). The customer service assistant using Alice can provide some basic customer services to customer at a customer service center (Demo: https://youtu.be/FsuORNP1kzQ).

The demo applications serve as proof of concept that the Alice framework can be customized for different applications. The combination of 3D avatar and awareness of identity and expression of the customer can provide personalized experience for the user.

**Acknowledgements** This project was financially funded by Telekom Malaysia Research and Development (TM R&D) Grant.

## References

1. Adam M, Wessel M, Benlian A (2020) AI-based chatbots in customer service and their effects on user compliance. Electron markets
2. Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: machines versus humans: the impact of artificial intelligence chatbot disclosure on customer purchases. Mark Sci 38(6):937–947
3. Herrera A, Yaguachi L, Piedra N (2019) Building conversational interface for customer support applied to open campus an open online course provider. In: 2019 IEEE 19th international conference on advanced learning technologies (ICALT), pp. 11–13
4. Patel NP, Parikh DR, Patel DA, Patel RR (2019) AI and web-based human-like interactive university chatbot (UNIBOT). In: 2019 3rd international conference on electronics, communication and aerospace technology (ICECA), pp. 148–150
5. Wu EH, Lin C, Ou Y, Liu C, Wang W, Chao C (2020) Advantages and constraints of a hybrid model K-12 e-learning assistant chatbot. IEEE Acc 8:77788–77801
6. Hoy MB (2018) Alexa, siri, cortana, and more: an introduction to voice assistants. Med Ref Ser Q 37(1):81–88
7. Këpuska V, Bohouta G (2018) Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In: 2018 IEEE 8th annual computing and communication workshop and conference (CCWC), pp. 99–103
8. Gelbukh A (2005) Natural language processing. In: Fifth international conference on hybrid intelligent systems (HIS'05). Rio de Janeiro, Brazil
9. Yu AW, Dohan D, Luong M, Zhao R, Chen K, Norouzi M, Le QV (2018) Qanet Combining local convolution with global self-attention for reading comprehension. In: Proc ICLR
10. Hofmann S, Reinecke M (2009) Cognitive–behavioral therapy with adults. Cambridge University Press
11. Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Mental Health 4(2)
12. Wan Y, Chiu C, Liang K, Chang P (2019) Midoriko chatbot: LSTM-based emotional 3D avatar. In: 2019 IEEE 8th global conference on consumer electronics (GCCE). Osaka, Japan, pp. 937–940

13. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
14. Angga PA, Fachri WE, Elevanita A, Suryadi, Agushinta RD (2015) Design of chatbot with 3D avatar, voice interface, and facial expression. In: 2015 international conference on science in information technology (ICSITech), pp. 326–330
15. Arsenijevic U, Jovic M (2019) Artificial intelligence marketing: chatbots. In: 2019 international conference on artificial intelligence: applications and innovations (IC-AIAI), pp. 19–193
16. Rhubarb Lip Sync. https://github.com/DanielSWolf/rhubarb-lip-sync
17. https://github.com/justadudewhohacks/face-api.js/
18. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758
19. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). Xi'an, pp. 67–74
20. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. https://arxiv.org/abs/1602.07360
21. Taniai H (2018), Keras-facenet. https://github.com/nyoki-mtl/keras-facenet
22. Serengil SI (2018) Deep face recognition with keras. https://sefiks.com/2018/08/06/deep-face-recognition-with-keras/
23. Knyazev B, Shvetsov R, Efremova N, Kuharenko A (2018) Leveraging large face recognition data for emotion classification. In: IEEE international conference on automatic face and gesture recognition (FG 2018). Xi'an, pp. 692–696
24. Kim JY, Liu C, Calvo RA, McCabe K, Taylor SCR, Schuller BW, Wu K (2019) A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. arXiv:1904.12403

# First Order Piecewise Collocation Solution of Fredholm Integral Equation Second Type Using SOR Iteration


**N. S. Mohamad, J. Sulaiman, A. Saudi, and N. F. A. Zainal**

**Abstract** We evaluate the first-order approximation solution piecewise by first-order polynomial collocation with Quadrature scheme on second-type Fredholm integral equations. This discretization derived the formulation to solve the first order piecewise approximation equation in which the linear system was built. The SOR method was described as a linear solver in which its formulation was constructed and applied in this study. In order to obtain the approximation solutions, the combination of SOR iterative method with the first-order piecewise polynomial by collocation with quadrature scheme has shown that performance of SOR method is superior than Jacobi method in terms of number of iterations and time of completion.

**Keywords** Polynomial · Collocation · Quadrature scheme · SOR · Jacobi · Fredholm integral equation of second type

## 1 Introduction

In recent findings, there are many authors who have chosen the integral equations of Fredholm as the analysis of their research because of the popularity of the integral equation topic among the mathematician. Not only that, integral equations have been commonly used in many areas and can be used in applied mathematics, chemistry,

N. S. Mohamad (✉) · J. Sulaiman · N. F. A. Zainal
Mathematics with Economics Program, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia
e-mail: norsyahida1302@gmail.com

J. Sulaiman
e-mail: jumat@ums.edu.my

N. F. A. Zainal
e-mail: farah.zainal19@gmail.com

A. Saudi
Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
e-mail: azali@ums.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_32

engineering, geophysics, electricity and magnetism. In this paper, we just pay attention at the second type Fredholm integral equation. Firstly, a simple briefing of the second type integral equation is the function of $U(x)$ that will be calculated while the function $g(t)$ is shown as the function given [1]. The parameter $\lambda$ plays a crucial role in obtaining the highest precision solutions and the number should be a non-zero values. The Fredholm integral equation (FIE) is divided into two types which is the linear Fredholm integral equation and non-linear Fredholm integral equation but in this paper, we only pay attention on the linear FIE only that was introduced by a few researchers [2, 3]. Basically, the Fredholm integral equation comes into many types which can be seen in some studies. The use of the Fredholm integral equation of first and third type has been widely used in this area of studies [4, 5]. Last but not least, the types of kernel must be thoroughly notified, as there are a few kernels types that have been implemented in the numerical calculation. The kernels that were found in previous studies such as smooth kernel, Hankel, weakly singular others [6].

Equations (1), (2) and (3) are the type of Fredholm integral equations (FIE) of one, two and three types that had been stated earlier.

$$g(x) = \int_a^b K(s,t)U(t)dt \tag{1}$$

$$U(x) + \lambda \int_a^b k(x,t)U(t)dt = g(t) \tag{2}$$

$$A(x)g(x) = U(x) + \int_a^b K(s,t)U(t)dt + P(x) \tag{3}$$

## 2 The Formulation of First-Order Piecewise Linear of Fredholm Integral Equation (FIE) of Second Type

In this paper, we show the finite grid of node points in details as we will have to derive the approximation equations by using the approach that was chosen. First of all, we need to focus on getting the approximation equation by using all node points. The first-order quadrature scheme, named the Trapezoidal scheme has also been involved in this process of discretization with the trapezoidal method. Later, we will discuss on how the trapezoidal scheme interferes in this formulation of approximation equations.

Figure 1 shows the function of Trapezoidal scheme which is introduced under the Newton-Cotes [7, 8]. The idea of implementation of the Trapezoidal scheme is used to create the approximation equation of Fredholm integral equation type two

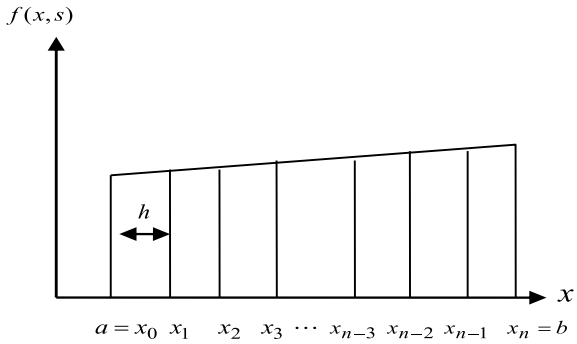



**Fig. 1** The function of trapezoidal scheme

through the discretization process of integral terms. The $h$ clearly shows the size of the sub-interval on the domain solution $[a, b]$.

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{(h)}{2}\left[f_i + f_{i+1}\right] \tag{4}$$

By referring to the Fig. 1, the function of Trapezoidal scheme on certain integration of function $f(x)$ on the interval of $[a, b]$ over the Eq. (4) can be seen as follow

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)dx$$

$$\int_a^b f(x)dx = \frac{(h)}{2}(f_0 + f_1) + \frac{(h)}{2}(f_1 + f_2) + \frac{(h)}{2}(f_2 + f_3) + \cdots + \frac{(h)}{2}(f_{n-1} + f_n)$$

$$\therefore \int_a^b f(x)dx = \frac{(h)}{2}\left(f_0 + 2\sum_{j=1}^{n-1} f_i + f_n\right) \tag{5}$$

We provide a simple overview in this section of the piecewise linear approximation equations in which this approximation equation is derived by considering the polynomial of the first order. In Eq. (2), where $\mathbf{U}(\mathbf{x})$, and $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{t})$ are defined as unknown function and kernel function respectively. Let us look at the solution domain [a, b] as shown in Fig. 2.

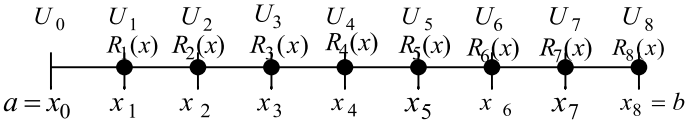



**Fig. 2** The finite grid with solution domain [a, b]

By referring to Fig. 2 with domain solution $\mathbf{I} = [\mathbf{a}, \mathbf{b}]$, the finite grid indicates that the integral concept is equipped in a number of situations. In the process of calculating the iteration numbers, the entire collocation node points were identified before they met the convergence criteria's [9]. The interval with $\mathbf{I} = [\mathbf{a}, \mathbf{b}]$ is used to describe the piecewise linear function. Equation (2) is a linear piecewise function that will be used to discrete and form the approximation equation of linear polynomial piecewise.

$$U(x) = \sum_{i=1}^{n} R_i(x).\delta_i(x) \tag{6}$$

Then, the approximation Eq. (7) formed as follows after the piecewise linear function was introduced in Eq. (2).

$$U(x) + \lambda \int_{x_0}^{x_1} k(x,t)R_1(t)dt + \lambda \int_{x_1}^{x_2} k(x,t)R_2(t)dt + \lambda \int_{x_2}^{x_3} k(x,t)R_3(t)dt + \ldots + \lambda \int_{x_n}^{x_{n-1}} k(x,t)R_n(t)dt = g(x) \tag{7}$$

After getting Eq. (7), the linear polynomial piecewise approximation equation will be modified and improvised with the implementation of collocation points into all of the node points of $x$ which is $x$ replaced with $x_i$. This is the highlight of the process which is the collocation points actually will be showing on all the finite domain. By considering all collocation points in the solution domain, the derivative of linear system can be formed.

The Eq. (7) can be rewrite as Eq. (8) after considering the collocation points into the approximation Fredholm integral equation of second type. Hence, the equation can be view as follow

$$U(x_i) + \lambda \int_{x_0}^{x_1} k(x_i,t)R_1(t)dt + \lambda \int_{x_1}^{x_2} k(x_i,t)R_2(t)dt + \lambda \int_{x_2}^{x_3} k(x_i,t)R_3(t)dt + \ldots + \lambda \int_{x_n}^{x_{n-1}} k(x_i,t)R_n(t)dt = g(x_i) \tag{8}$$

Referring to all collocation points, the first-order piecewise approximation equations can be considered as

$$G\underline{U} = \underline{g} \tag{9}$$

where

$$G = \begin{bmatrix} 1 + G_n(x_n) & \cdots & G_n(x_n) \\ \vdots & \ddots & \vdots \\ G_n(x_n) & \cdots & 1 + G_n(x_n) \end{bmatrix}, n = 0, 1, 3, \ldots, n$$

$$\underline{U} = [U_0, \ldots, U_n],^T n = 0, 1, 3, \ldots, n$$

$$\underline{g} = [g_0, \ldots, g_n],^T n = 0, 1, 3, \ldots, n$$

## 3 Illustrative Example and Iterative Method

In order to obtain the result of the first-order piecewise polynomial solution of Fredholm integral equation, there are several iterative methods that we can use such as Jacobi, Gauss-Seidel and SOR iterative method. These iterative methods are used to determine the approximation solutions of system linear. Thus, we can make the comparison of the performance of each iterative method together with the approximation equation that was created. The iteration solver of this study is the SOR iterative method. We will make a quick review about the SOR iterative method later [10, 11]. It was created based on the modification of Gauss-Seidel iterative method which has slightly difference in the algorithm which it carries the weighted parameter. It helps in getting the highest convergence test [12–14]. The general formulation of SOR method can be stated as

$$\underline{U}^{k+1} = (1 - w)\underline{U}^k + w(D + C)^{-1}\left(-V\underline{U}^{(k)} + \underline{g}\right) \tag{10}$$

Referring to Eq. (10), the matrices D + C + V are portraying the diagonal, upper and lower matrices respectively. The implementation of this SOR method can be illustrated in Algorithm 1 and indicated in Fig. 3.

**Algorithm 1** *SOR iterative method*

1. State the initial value of $\underline{U}^0 \leftarrow 0, \varepsilon \leftarrow 10^{-10}$;
2. Calculate $i = 1, 2, 3, \ldots, n, \underline{U}^{(k+1)} = (D + C)^{-1}\left(V\underline{U}^{(k)} + \underline{g}\right)$
3. Check the convergence, $\left|U_i^{(k+1)} - U_i^{(k)}\right| \leq \varepsilon = 10^{-10}$. If yes go to step iv. Otherwise, repeat the step ii.

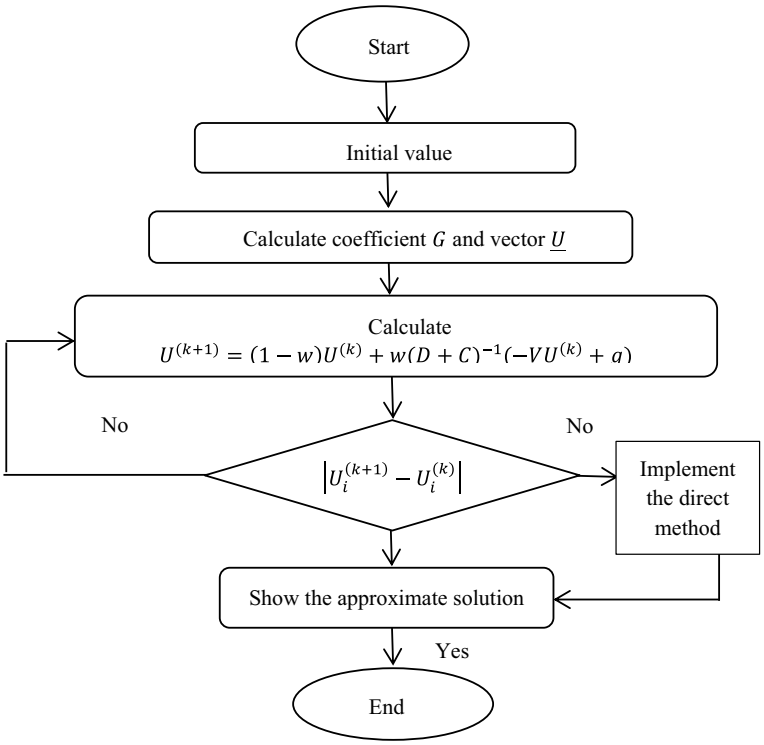



**Fig. 3** The lane of the flowchart of SOR implementation

4. Show the approximate solutions.

To test the performance of the presented iterative methods, there are three examples of Fredholm integral equation of type two that carried out on approximation equation of Fredholm integral of type two. Three numerical parameters have been highlighted such as the iteration numbers (L), time of completion (s) and the Max. Abs. Error. We tested five different sizes of $n$ which are 512, 1024, 2048, 4096 and 8192. The numerical analysis is tested on the following three examples of Fredholm integral equation of second type in which, we will see them as follows:

***Example 1*** This example was taken [15].

$$U(x) = e^{-xi} - \frac{1}{2} + \frac{1}{2}e^{(x+1)} + \frac{1}{2}\int_0^1 (x+1)e^{-xy}U(t)dt \tag{11}$$

while having exact solution is

$$U(x) = e^{-x} \tag{12}$$

***Example 2*** The following is Fredholm linear of second type integral equation [16].

$$U(x) = e^{3x} - \frac{1}{9}\left(2e^3 + 1\right)x + \int_0^1 xtU(t)dt, 0 < x < 1 \tag{13}$$

With exact solution is given by

$$U(x) = e^{3x} \tag{14}$$

***Example 3*** Let the second type of the linear Fredholm integral equation be mentioned as [17].

$$U(x) = e^x - 1 + \int_0^1 tU(t)dt \tag{15}$$

With exact solution is stated as

$$U(x) = e^x \tag{16}$$

Based on the Tables 1, 2, 3 and 4 and Figs. 3, 4, 5, 6, 7, 8 and 9, the iterative methods have been tested with the combination of Trapezoidal scheme with first-order piecewise polynomial approximation equation of Fredholm integral equation

**Table 1** The distinction of between number of iterations based on Jacobi, GS and SOR iterative method

| Ex. | n | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|
| | Method | Iteration numbers (L) | | | | |
| 1 | Jacobi | 41 | 41 | 41 | 41 | 41 |
| | GS | 23 | 24 | 24 | 24 | 24 |
| | SOR | 14 | 14 | 14 | 14 | 14 |
| 2 | Jacobi | 24 | 24 | 24 | 24 | 24 |
| | GS | 15 | 15 | 15 | 15 | 15 |
| | SOR | 11 | 11 | 11 | 11 | 11 |
| 3 | Jacobi | 35 | 35 | 35 | 35 | 35 |
| | GS | 21 | 21 | 21 | 21 | 21 |
| | SOR | 13 | 13 | 13 | 13 | 13 |

**Table 2** The distinction of time of completion based on Jacobi, GS and SOR iterative method

| Ex. | n | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|
| | Method | Time of completion (s) | | | | |
| 1 | Jacobi | 16.58 | 66.16 | 264.63 | 1056.41 | 4220.96 |
| | GS | 9.45 | 38.65 | 154.65 | 620.67 | 2474.18 |
| | SOR | 5.92 | 22.78 | 90.79 | 361.47 | 1450.66 |
| 2 | Jacobi | 1.37 | 5.44 | 2.33 | 84.63 | 339.47 |
| | GS | 0.93 | 3.59 | 14.5 | 56.48 | 211.47 |
| | SOR | 0.67 | 2.58 | 9.91 | 39.16 | 156.19 |
| 3 | Jacobi | 2.05 | 8.16 | 32.01 | 127.61 | 510.88 |
| | GS | 1.25 | 4.83 | 19.25 | 76.8 | 306.5 |
| | SOR | 0.83 | 3.13 | 12.01 | 47.87 | 190.27 |

**Table 3** The distinction of Max. Abs. Error based on Jacobi, GS and SOR iterative method

| Ex. | n | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|
| | Method | Max. Abs. Error | | | | |
| 1 | Jacobi | 3.25E-06 | 3.25E-06 | 3.25E-06 | 3.25E-06 | 3.25E-06 |
| | GS | 3.25E-06 | 3.25E-06 | 3.25E-06 | 3.25E-06 | 3.25E-06 |
| | SOR | 3.25E-06 | 3.25E-06 | 3.25E-06 | 3.25E-06 | 3.25E-06 |
| 2 | Jacobi | 1.69E-05 | 1.69E-05 | 1.69E-05 | 1.69E-05 | 1.69E-05 |
| | GS | 1.69E-05 | 1.69E-05 | 1.69E-05 | 1.69E-05 | 1.69E-05 |
| | SOR | 1.69E-05 | 1.69E-05 | 1.69E-05 | 1.69E-05 | 1.69E-05 |
| 3 | Jacobi | 3.16E-06 | 3.16E-06 | 3.16E-06 | 3.16E-06 | 3.16E-06 |
| | GS | 3.16E-06 | 3.16E-06 | 3.16E-06 | 3.16E-06 | 3.16E-06 |
| | SOR | 7.91E-07 | 7.91E-07 | 7.91E-07 | 7.91E-07 | 7.91E-07 |

**Table 4** Reduction percentage of iteration numbers and time of completion for Jacobi, GS and SOR iterative method on three examples

| Example | Iteration numbers (L) (%) | Time (s) (%) |
|---|---|---|
| 1 | 39.13–41.66 | 21.29–24.53 |
| 2 | 26.66 | 16.28–21.51 |
| 3 | 38.09 | 20.48–22.75 |

of second type. By referring to the Tables 1 and 2, we have discussed in context of iteration numbers and time completion. As we can see, the iteration numbers and time of completion of SOR iterative method is smaller than Jacobi and Gauss-Seidel method. The percentage reduction of iteration numbers also shows a big number of percentage reduction with 39.13–41.66%, 26.66% and 38.09% respectively. While the percentage reduction of time of completion are 21.29–24.53%, 16.28–21.51% and 20.48–22.75% respectively. The accuracy of the Max. Abs. Error also increasing

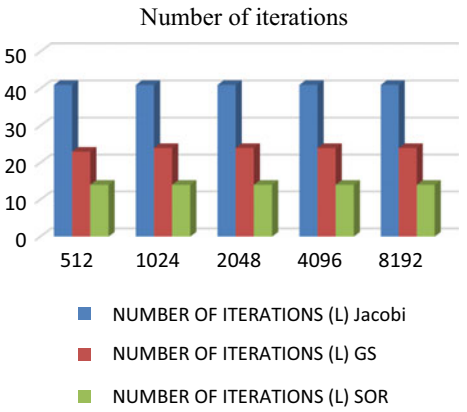



**Fig. 4** The comparison of iteration numbers of Jacobi, GS and SOR for example 1

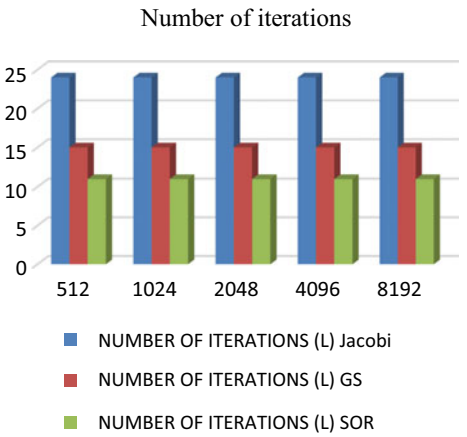



**Fig. 5** The comparison of iteration numbers of Jacobi, GS and SOR for example 2

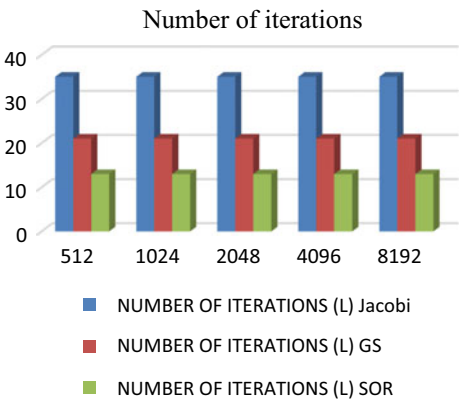



**Fig. 6** The comparison of iteration numbers of Jacobi, GS and SOR for example 3

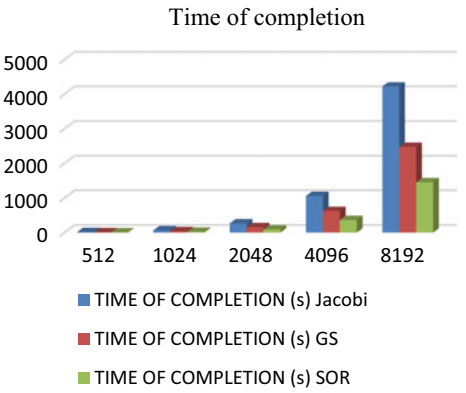



**Fig. 7** The comparison time of completion of Jacobi, GS and SOR for example 1

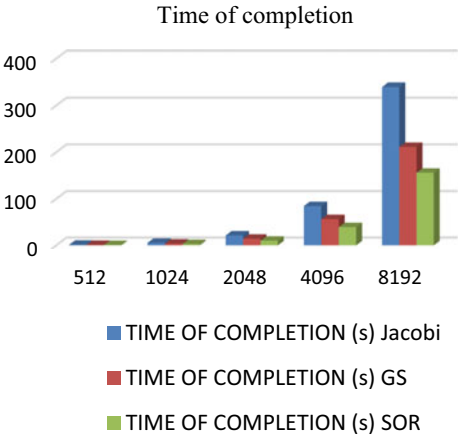



**Fig. 8** The comparison time of completion of Jacobi, GS and SOR for example 2

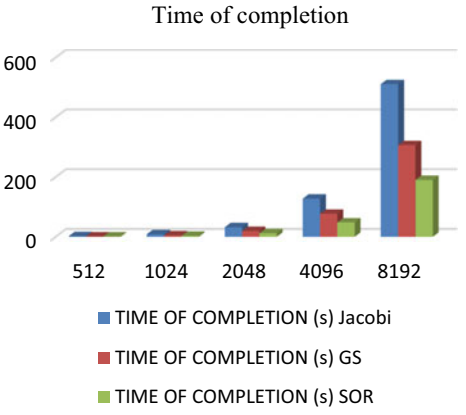



**Fig. 9** The comparison time of completion of Jacobi, GS and SOR for example 3

with respect to $n$ based on Table 3. Clearly, the number iteration of Jacobi is the highest due to the main characteristic of Jacobi method always uses the previous value in its formulation. While, other method considers the latest values in the formulation during the iteration process.

## 4 Conclusion

In this paper, the piecewise linear function with colocation approach was introduced in the previous section in order to create a linear system from the discretization process of the linear second kind Fredholm integral equation. The iterative SOR method has also decreased its number of iterations and time of completion with which the numerical results were displayed. As a conclusion, the SOR iterative method output in conjunction with the combination of piecewise linear function and collocation method is better than Jacobi and Gauss-Seidel method.

**Acknowledgements** The author would like to thank Postgraduate Centre, Universiti Malaysia Sabah for funding this paper.

## References

1. Mohamad NS, Sulaiman J (2018) The piecewise polynomial collocation method for the solution of Fredholm equation of second kind by using age iteration. J Phy:012039
2. Ahmed A (2016) Numerical solution of linear and non-linear Fredholm integral equations by using weighted mean-value theorem. Springer Plus 5:1962
3. Esmaeili H, Moazami D (2019) A kernel-based technique to solve three-dimensional linear Fredholm integral equations of the second kind over general domains. Comput Appl Math 19(8):181
4. Gabbasov NS (2019) Special version of the spline method for integral equations of the third kind. Diff Equ 55(9):1261–1268
5. Mesgarani H, Azari Y (2019) Numerical integration of Fredholm integral equations of the first kind with noisy data. Math Sci 13:267–278
6. Mohamad NS (2018) The piecewise collocation solution of second kind Fredholm integral equations by using quarter-sweep iteration. J Phy 1358:012052
7. Moheuddin MD, Uddin MJ, Kowsher M (2019) A new study of trapezoidal, simpson's 1/3 and simpson's 3/8 rules of numerical integral problems. Appl Math Sci 6(4)
8. Uysal U, Taymaz I (2019) Experimental investigation of heat transfer on trapezoidal channel with three passes. J Therm Anal Calorim 140:953–964
9. Srrivastava S, Stanimirovi'c PS, Katsikis VN, Gupta KD (2017) A family of iterative methods with accelerated convergence for restricted linear system of equations. Mediterr J Math 14:222
10. Mohamad NS, Sulaiman J (2018) The piecewise polynomial collocation method for the solution of Fredholm equation of second kind by using SOR iteration. In: AIP conference proceeding, vol. 2013, pp. 020037
11. Grzegorski SM (2019) On optimal parameter not only for the SOR method. Appl Comput Math 8(5):82–87
12. Saha M, Chakrabarty J (2018) On generalized jacobi, gauss-seidel and SOR methods

13. Sathya S, Ramesh T (2019) Comparison of gauss jacobi method and gauss-seidel method using scilab. In: International Journal of Trend in Scientific Research and Development, pp. 2456–6470
14. Akram M, Muhammad G, Koam ANA, Hussain N (2019) Iterative methods for solving a system of linear equations in a bipolar fuzzy environment. https://www.mdpi.com/journal/mathematics 7(8):728
15. Rodwellhead (2014) https://www.scribd.com/document/214615700/Solution-of-Fredholm-Integral-by-Collocation. Accessed 2014
16. Paradin N, Gholomtabar Sh (2010) Numerical solution of the linear Fredholm integral equations of the second kind. J Math Extension 5(1):31–39
17. Avazzadeh Z, Heydari M, Loghmani GB (2010) Numerical solution of Fredholm integral equations of the second kind by using integral mean value theorem. Appl Math Modell 35:2374–2383

# Vision-Based Activity Recognition System with a Deep Neural Network for Surveillance


**Suheib Faisal Abubaker Sherif, Ooi Chee Pun, Tan Wooi Haw, and Tan Yi Fei**

**Abstract** Computer vision has gained tremendous attention recently due to what visual data can provide in terms of meaningful information and predictions. Videos, not like other types of data, can carry lots of information about the captured scene. Information such as objects detection, face recognition and action classification can be beneficial to monitoring systems such as traffic monitoring and security systems. Activities recognition, in particular, is a quite significant part of visual data analysis and can provide pragmatic predictions on people's behaviour. The absence of a well-labelled video dataset makes it more challenging to develop machine learning algorithms for irregular actions recognition. These prediction models can ease the process of monitoring buildings, roads and other common areas monitored by CCTV systems. This paper proposes a method to utilise deep learning in classifying people's behavior by identifying normal behaviours and classifying any unusual activities and provide a well-trimmed and labelled dataset for abnormal behaviors in CCTV videos.

**Keywords** Computer vision · Deep learning · Activity recognition

## 1 Introduction

### *1.1 Background*

Detecting unusual behaviours in security footage and live feed is currently done manually by relying on security guards to monitor the screens all the time. This process is tedious, especially when many angles need to be monitored continuously and requires a human observer to operate. Visual data obtained from videos provide

S. F. A. Sherif (✉) · O. C. Pun · T. W. Haw · T. Y. Fei
Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
e-mail: suheib.sherif@gmail.com

O. C. Pun
e-mail: cpooi@mmu.edu.my

T. W. Haw
e-mail: twhaw@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_33

plenty of information on the status of the building or street along with many other statistics on actions and behaviour of the objects within the video. By analysing this information, the process of monitoring security cameras can be automated to include behaviour and status predictions.

Artificial intelligence is a popular topic for what it can achieve in systems automation and data analytic. The main challenge that many researchers face in this step is finding a well-labelled dataset to train a machine learning model for action classification and prediction. Many researchers focused on identifying people positions or activities such as sitting, walking, running, etc. Surveillance systems, nowadays, still rely on human observers to detect irregular behaviour that is happening in the building or street. This approach required not only observers to monitor multiple screens simultaneously but also costly and prone to human errors. With the advancement of technology, this process can be automated for irregular actions detection.

## *1.2 Literature Review*

Computer vision has been an important topic for many researchers. Many efforts were spent to identify visual information included in images and videos. Some researchers focused on understanding human action behaviour and body movements while others concentrate on understanding the relationship between objects and people inside the frame or video. Yalçın et al. [1] introduced the classification of human activities using deep learning by extracting human's joint coordination during the actions. While avoiding manual feature extraction for a better result, their method relies on specialised hardware to detect human skeleton and can only identify precise subjects in the frame. Babiker et al. [2] introduced a system that can recognise human activities by using a set of image processing techniques. Their process mainly involved removing background from videos followed by noise reduction and then extracting the features such as body bounding box, the centre of gravity and the location of the body. These features were then used to train a perception network model for classification. However, they were only able to detect human activities such as walking, sitting etc. Murad et al. [3] adopted deep recurrent neural networks in human activities recognition using data extracted from gyroscopes and accelerometers. They utilised Long Short-Term Memory to capture the temporal changes in readings from the sensors. Bhardwaj et al. [4] used K-means clustering to categorise the subjects in the video and use the clusters to train a neural network to detect body positions such as standing, jumping, etc. Another research by Mo et al. [5] utilised convolutional neural networks for feature extraction and a multi-layer perception model for video classification. They used Kinect camera to extract the skeleton structure used in model training. Other researchers focused on studying group relations in videos, Lee et al. [6] was able to detect the behaviour of people walking along the street such as people walking in a group or splitting apart, etc. They used the number of people, velocity and position as fixed features to train the model.

Sun et al. [7] highlighted that the most critical parameter to improve the model accuracy is to have more frames in the training process. However, that does not guarantee better results every time. Performance can be affected by the duration of the video frames or the time range [7]. Ng et al. [8] proposed another method of video classification by exploring optical flow data obtained from the video frames. Their approach is to take videos up to two minutes and extract optical flow as one shot, then make use of this information in training and predictions. Other researchers focused on optical flow extraction, and they were able to obtain excellent results. However, their method can be useful in classifying videos as a single process, which is not suitable for continuous classification of video frames [8, 9].

All these literature reviews that, to detect anomaly in videos or classify videos based on their actions, the study of the temporal behaviour of the subjects within the video [10, 11] is essential. That means the type of action can only be identified by extracting the features from a sequence of video frames.

### *1.3 Motivation*

The main objective of this project is to automate the monitoring process of surveillance systems by developing a system that can predict the actions and behaviours happening in these videos. With the widely used of CCTV systems, vast amounts of videos are being streamed and recorded every day. Thus, providing a tremendous opportunity to analyse and fetch out meaningful information from these videos which can produce powerful prediction tools in all fields. Hence, in this project, the machine learning approach is employed to predict and identify the irregular behaviour in the surveillance systems.

## 2 System Architecture and Data Analysis

The system, (see Fig. 1), is comprised of three main segments: dataset processing, machine learning model and the prediction process on human behaviour.

### *2.1 Dataset Preparation*

The main concept of training a machine learning models is to have a well-labelled dataset containing relevant features along with a class or label for each element. Then, this dataset can be utilised to train a model to predict the label based on a given set of features. The first step towards preparing the dataset is to group each category of videos needed to predict in one directory and name it according to its action name. The dataset used in this project is taken from different sources of CCTV footage

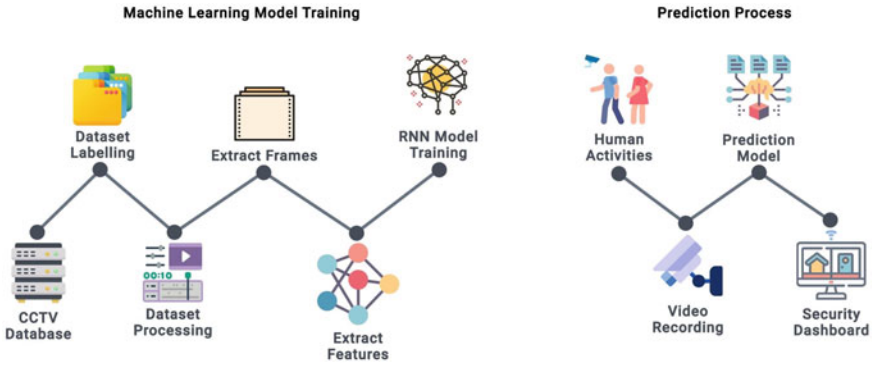



**Fig. 1** System architecture

uploaded to the internet, including UCF-Crime dataset and YouTube videos [11]. After getting the videos grouped and labelled correctly, further processing is needed to increase the accuracy of the training. The second process is to trim unwanted data from videos, which means only the section of the video containing the action is kept. The repeated frames with an idle screen will be removed, which is common in many CCTV footage. This process is automated by checking the changes in the video frames and trimming the repeated frames. Then, a manual process is carried out to obtain accurate labelling of the dataset with specific action; That means, if one video under the vandalism category, it should only show the vandalism action in the video, no normal behaviours should appear in the video. This labelling process is done manually by watching the scenes happening in each video and trimming the video accordingly. This process has cut the dataset size by around 35% from the original set, which provides more accurate data for training.

## *2.2 Training Recurrent Neural Network Model*

It remains a challenge to understand the context of the videos with a single video frame. For example, looking at one person standing in front of a door in one frame might seem normal, but adding the information of this person hitting the door in the following frames will indicate that some sort of vandalism or break-in is happening. Thus, the temporal changes in video frames extracted is essential. One of the most popular techniques in Recurrent Neural Networks is Long Short-Term Memory—LSTM. Unlike the usual neural networks, LSTM has a feedback connection where predictions are not only made based on the current inputs but rather take into account the outputs from previous nodes as well [12, 13]. This technique will allow us to consider the targeted temporal behaviour extracted from the video frames. In this paper, the semi-supervised training is adopted where only the labelled dataset and specific videos frames were used to train the model. The LSTM model uses the

**Table 1** TensorFlow sequential model design parameters

| Layer (type) | Output shape |
|---|---|
| lstm (LSTM) | (None, 2048) |
| dence (dence) | (None, 512) |
| dropout (droupout) | (None, 512) |
| dence_1 (dence) | (None, 3) |

features extracted from the widely known Convolutional Neural Networks models and feed the features extracted to train the model. We first pass the frame to image classification model such as Inception Network; then we move the features layer to the input of our model. Dataset is divided into training and testing folders with 80% training and 20% testing. The training was done using Keras with TensorFlow back-end and promising prediction results is generated [14].

Table 1 presents the design flow of the Recurrent Neural Network; the LSTM layer takes input from the extracted features. The feature extraction process is done by saving the features extracted by InceptionV3 model at the final pool layer. Next, the input is passed to a 2048 nodes LSTM layer followed by a 512 nodes dense layer then a dropout layer of 512 nodes and finally a dense layer with 3 nodes for the output. The model was configured with Keras Adam optimiser with $1 \times 10^{-5}$ learning rate.

### *2.3 Application Layer*

After training the video classification model, and to use this model, an application layer is needed. Since our model does not take raw videos as input to make predictions, we need a software to extract the sequences of frames from videos, extract features from these frames using ConvNets and pass the extracted features to the trained model to make the prediction. Thus, the process of the prediction is not real-time, but a slight delay is expected where frames are extracted and processed before making the predictions.

## 3 Results and Discussion

### *3.1 Dataset Pre-processing*

Labelling and grouping videos in different categories is a necessary step before training the model. Since the structure of the project is fixed, it is easier to add in more prediction categories and retrain the model. The pre-processing of data is a key element here, by providing well-trimmed videos and accurate labelling,

a better model can be trained. The dataset size was reduced by 35% through the removal of unwanted scenes from videos and removing repeated frames which do not carry extra information; it also reduced training time and process as well. The initial training of the model included two anomaly categories; each category included 50 videos trimmed to contain the action only. The two categories are 'Fighting' and 'Road Accidents' which included some videos of people fighting along the streets where cars are passing by. Such a condition will test the trained model accuracy and limitations. The videos were randomly divided into 80% for training and 20% for validation.

## *3.2 Model Training*

The first step towards training the model was to extract frames from the videos and store the frame sequences in a different directory. This process was done by using FFmpeg, which is an open-source software to edit videos, images and other multimedia files. After extracting image frames from videos, each frame is passed through a classification model for feature extraction. The model selected in this stage is Inception Model V3 which is widely used in image classification. By using Keras, a widely known Python library for machine learning that runs on TensorFlow open-source library, an LSTM model was trained with the extracted features. The produced model can then identify a video category given a set of frames sequence. The methodology was initially tested with two types of actions; each action included only 30 videos along with a normal videos category that included videos of people interacting, cars passing by and other normal actions and it showed promising results. Further training will be done using the complete dataset and validated using a new set of videos.

Tables 2 and 3 presents the count of parameters fetched by TensorFlow for training. The training process was set to 20 Epochs of training with early stopping if the

**Table 2** TensorFlow sequential model design parameters

| Layer (type) | Output shape | Parameters count |
|---|---|---|
| lstm (LSTM) | (None, 2048) | 33,562,624 |
| dense (dense) | (None, 512) | 1,049,088 |
| dropout (dropout) | (None, 512) | 0 |
| dence_1 (dense) | (None, 3) | 1539 |

**Table 3** TensorFlow training parameters

| | |
|---|---|
| Total params | 34,613,251 |
| Trainable params | 34,613,251 |
| Non-trainable params | 0 |

**Table 4** LSTM training and validation accuracy

| Name | Loss | Steps | Duration |
|---|---|---|---|
| LSTM/train | 0.2905 | 15 | 5 m 52 s |
| LSTM/validation | 0.1239 | 15 | 5 m 52 s |

resulting model is not improving further. The training was completed in 16 Epochs and the validation loss improved from 0.493 to 0.1239, as shown in Table 4.

As shown in Fig. 2, the validation accuracy improved from 0.652 to 0.9756. Figure 3 represents the improvement of validation accuracy and validation loss during the training process, which was generated by TensorBoard [14] (Table 5).

The loss and accuracy of the LSTM model were compared to the Multilayer Perceptron (MLP) model, and it showed better results. The results obtained from


**Fig. 2** LSTM training and validation accuracy


**Fig. 3** LSTM training and validation loss

**Table 5** LSTM training and validation loss

| Name | Loss | Steps | Duration |
|---|---|---|---|
| LSTM/train | 0.875 | 15 | 5 m 52 s |
| LSTM/validation | 0.9756 | 15 | 5 m 52 s |

the initial training was satisfactory, and further dataset processing and more action categories will be added for future model improvements. The LSTM model is trained using different versions of the dataset; It is observed that the more frames included in the training sequence, the better the results [15]. Hence 450 frames from each video are used in order to have sufficient information from the video and yet not too complicated by not setting a limit for the training sequence [12]. Videos were between 4 and 15 s containing the labelled action only. Another parameter that affects the performance of the LSTM is the number of epochs of training, and the training process was set to 30 epochs with an automatic stop if the model is no longer improving in terms of validation loss

### *3.3 Dashboard Monitoring*

In order to automate the process of monitoring security footage from live cameras, or detect anomaly in recorded videos, a software tool is needed to process these videos and predict the action happening at each scene. This software is developed using Python and FFmpeg to extract frames then apply classification model for feature extraction then get these features and pass them as input to the trained model. The predicted category with the highest probability will later be displayed as text label while playing the video.

## 4 Conclusion and Future Direction

This project aims to automate the behaviour monitoring in surveillance systems by using artificial intelligence. Videos were grouped and labelled according to the type of action happening in the video. Then, unwanted scenes were trimmed from the videos so we can get a more accurate dataset for training. A widely known image classification model, Inception Network, was used for feature extraction from video frames. These features were then fed to train an LSTM model which consider the temporal behaviour happening in videos that indicates the type of action. The method was tested with a subset of the dataset, and promising results are observed. The primary constraint in this project is a lack of large dataset for normal and abnormal actions that is precisely trimmed to contain the labelled action. Further improvements and complete dataset training will be conducted and verified in future. Improvements will be made toward the dataset preparation and the application layer built on top of the trained model.

## References


1. Yalçın M, Tüfek N, Yalcin H (2018) Activity recognition of interacting people. In: 2018 26th signal processing and communications applications conference (SIU). IEEE, pp 1–4
2. Babiker M, Khalifa OO, Htike KK, Hassan A, Zaharadeen M (2017) Automated daily human activity recognition for video surveillance using neural network. In: 2017 IEEE 4th international conference on smart instrumentation, measurement and application (ICSIMA). IEEE, pp 1–5
3. Murad A, Pyun JY (2017) Deep recurrent neural networks for human activity recognition. Sensors 17(11):2556
4. Bhardwaj R, Kumar S, Gupta SC (2017) Human activity recognition in real world. In: 2017 2nd international conference on telecommunication and networks (TEL-NET). IEEE, pp. 1–6
5. Mo L, Li F, Zhu Y, Huang A (2016) Human physical activity recognition based on computer vision with deep learning model. In: 2016 IEEE international instrumentation and measurement technology conference proceedings. IEEE, pp 1–6
6. Lee DG, Kim PS, Lee SW (2017) Local group relationship analysis for group activity recognition. In: 2017 17th international conference on control, automation and systems (ICCAS). IEEE, pp 236–238
7. Sun J, Wang J, Yeh TC (2017) Video understanding: from video classification to captioning. In: Computer vision and pattern recognition. Stanford University, pp 1–9
8. Ng JY-H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
9. Zhu L, Yang Y (2020) Label independent memory for semi-supervised few-shot video classification. IEEE Trans Pattern Anal Mach Intell
10. Landi F, Snoek CG, Cucchiara R (2019) Anomaly locality in video surveillance. arXiv preprint arXiv:1901.10364
11. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6479–6488
12. Tang Y, Yang J, Chen J, Collge X Original paper comparative research on influencing factors of LSTM deep neural network in stock market time series prediction
13. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
14. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, … Ghemawat S (2015) TensorFlow: large-scale machine learning on heterogeneous systems
15. Boulmaiz T, Guermoui M, Boutaghane H Impact of training data size on the LSTM performances for rainfall–runoff modeling

# A Scalable Cloud-Based Medical Adherence System with Data Analytic for Enabling Home Hospitalization


**Abubaker Faisal Abubaker Sherif, Tan Wooi Haw, Ooi Chee Pun, and Tan Yi Fei**

**Abstract** Medication non-adherence is one of the most significant concerns in managing chronic diseases which has inevitable consequences. While various technologies and research have been developed and carried out to monitor medical adherence for patients, their approaches lack in terms of the assurance of medicine consumption and the cost effectiveness of their solutions. This paper provides a cloud-based medical adherence system that can track patients' medicine intake based on the physical effects of the medicine on their bodies by tracking their vital signs. A machine learning model is trained to classify the patient health status and this data is used to determine whether their bodies are responding to the medicine, which is used to alert doctors to enable home hospitalization. The use of this system is proposed to serve as a secondary decision support provider to compliment and ease the decision-making process done by doctors.

**Keywords** Medical adherence · Home hospitalization · Machine learning

## 1 Introduction

### *1.1 Background*

Medicine adherence is defined as the extent to which a patient follows the treatment plan by consuming the medicines prescribed by their doctor [1]. Not-adhering to a treatment plan can result in undesirable clinical consequences and substantial increase in hospitalization costs [2]. Recent research highlighted that most patients

A. F. A. Sherif (✉) · T. W. Haw · O. C. Pun · T. Y. Fei
Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
e-mail: abubaker_fs@yahoo.com

T. W. Haw
e-mail: twhaw@mmu.edu.my

O. C. Pun
e-mail: cpooi@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_34

do not follow the instructions provided by their doctors which results in significant increase in hospitalizations and doctor visits [3].

In recent years, home hospitalization technologies have been the focus of many researchers and healthcare companies [4]. The two main aspects of home hospitalization are medicine adherence and health status monitoring for patients. Different approaches to detect medical adherence are used in this field, some use blood test evaluation, some rely on patient's self-assessment feedback while others use pill dispensers to detect whether the medicine was taken. These solutions can provide useful information on the behavior of the patient, but they lack in terms of cost efficiency, illness type and how practical they can be achieved.

This paper aims to provide a solution that can assist doctors to monitor the health status and medicine intake by analyzing the patient's vital signs throughout the day. We leverage machine learning to identify the health condition of the patient and match that to the timeline set by doctors for medicine intake. The trained model takes major vital signs as input which are measured by off-the-shelf medical proven instruments. The data taken is logged and processed on cloud and provides the output in terms of recommendation to doctors who can then decide to adjust medicine dosage or frequency or schedule an appointment with the patient.

## *1.2 Literature Review*

Many efforts were made in the field of medical adherence and home hospitalization. Providing a home hospitalization solution to patients proved to reduce in-hospital days of patients [5]. One key feature that is necessary for enabling home hospitalization is the ability to monitor the patient's health status remotely and track their adherence to medications. Tripathi et al. [6] proposed the usage of tracking sensors and wearable devices to collect information about the patient, then the data is pushed to an online server where the decision is made to contact family members, ambulance or clinical support. Other researchers such as Zulkifli et al. [7] developed a health monitoring and information system to enable smart health environment. Their solution links patients, health service staff and doctors. Patients can provide feedback to doctors using their smartphones and doctors will respond to their reports to check if they need to schedule an appointment for the patient or continue treatment from home. Home hospitalization is receiving huge attention for what it can provide in terms of reduction of hospitalization time, reduction of treatment cost and freeing up doctors' time. Hernández et al. [8] conducted a study on early discharge care service by enabling patients to stay at home for some time of their hospitalization treatment. They requested nurses to visit these patients frequently and measure their vital signs and enter them to an online system for doctors to access. The study was conducted for ten years and provided promising results as the service reduced the hospitalization time by six days [8]. Lobato et al. [9] conducted a study on forty patients divided into two groups, the first group received their care at the hospital while the other group was sent back home and had their health status monitored regularly and received

the same treatment as the first group. After the patients fully recovered and were discharged from the hospital, it was noticed that there is no significant difference between patients who were staying at home and those who stayed at the hospital. Federman et al. [10] conducted a study on patients who require inpatient-care and sent a group of them to receive care at home. Patients had their vital signs monitored by nurses and health professionals, then all patients were asked to review their treatment process and the results showed better rating for those who received treatment at home. Sherif et al. [11] proposed a method to enable home hospitalization by tracking medicine intake for patients using embedded hardware. Their approach relied on patients reporting their medicine adherence using an alarm button which connects to a monitoring dashboard. This approach provides a good method for patients monitoring but it does not guarantee that the patient has taken the medicine when they send the acknowledgment. This can only be verified by tracking patient physical condition and monitoring their vital signs. Daramola et al. [12] followed a similar method by reporting medicine intake using smartphone application and rely on patient self-reporting which does not guarantee medicine intake. Kumar et al. [13] proposed a similar approach for medicine adherence tracking using medicine dispenser. This method assumes that the medicine was taken whenever the patient opens the dispenser to take medicine. However, this also does not guarantee the medicine intake. Hence, this paper aims to verify the consumption of medicine by tracking vital signs of patients and reporting irregular readings regularly.

### *1.3 Motivation*

The primary objective of this study is to enhance patients' adherence to medication without the need of special equipment or the supervision of nurses and caregivers. With the recent spread of COVID-19 pandemic, where thousands of people have been hospitalized and many healthcare facilities ran out of resources to accommodate the increasing numbers of patients, many patients with minor symptoms were asked to stay at home and monitor their health status [14]. This pandemic raised the significant need to enable home hospitalization. Thus, we propose a complimentary tool that helps to predict the behavior of the patients and highlight their behavior to their doctors so they can make better decisions while patients rest at home.

## 2 System Architecture and Data Analysis

The proposed method composes of three main sections: data preparation and data preprocessing, training and validating a machine learning model to classify medication adherence behavior and a method for doctors to validate the result and advise on the patient condition based on the given prediction as well as data collected from patients.

## 2.1 Data Pre-processing and Data Collection

The four main indicators of health status for patients are heart rate, blood pressure, temperature and blood oxygen saturation. Thus, these parameters are taken into consideration in our design where the focus was to prepare a labeled dataset that is collected by medical professionals for patients diagnosed with medical conditions. We leveraged on a renown public health dataset known as Medical Information Mart for Intensive Care III—MIMIC-III which includes health-related data of more than 40,000 patients who were admitted to intensive care units—ICU. The dataset contains abundant information about patients including readings of vital signs as well as doctors diagnosis. Data were taken at 1-h frequency and keyed into the system by caregivers. However, the data is not consistent in terms of what data is available for each patient. In order to train an accurate machine learning model, the data must be properly formatted. We filtered out the patients who had more than four readings taken per day including heart rate, blood pressure, temperature and oxygen saturation along with doctor diagnosis.

Machine learning models for classification need data for normal people, not diagnosed with any medical condition as well as the prepared data for patients diagnosed with diseases. However, there is no publicly available dataset with the criteria that we have, so we collected data for healthy people to be used for the model training.

## 2.2 Training Recurrent Neural Network Model

Two models are proposed to provide useful information as a second opinion for doctors. The first model is a binary classifier to identify patient health status as normal or abnormal depending on the vital signs measured. The prediction will be based on the labeled dataset provided with normal and abnormal flags.

The binary classifier aims to provide a prediction based on the vital signs labeled data by grouping all diagnoses under one single category labeled as abnormal readings while normal readings are labeled as normal.

Three types of classifiers are selected to be tested and validated with the prepared dataset. These classifiers are K-NN, linear Support Vector Machine (SVM) and kernelized SVM with radial basis function (RBF). The models are selected according to the size of the dataset that we prepared and the type of classification that is needed. Model training will be using Scikit-learn which is well known for classification algorithms [15].

The multiclass classifier aims to provide a predictive suggestion on the diagnostic category of the patient based on the vital signs. Diagnoses are categorized into different groups such as heart issues, respiratory illness, blood pressure etc., which the doctor will later confirm or deny to provide a feedback data that can be used later for training a more accurate model.

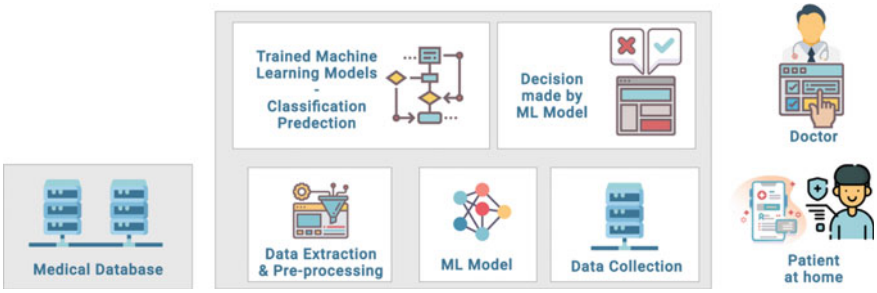



**Fig. 1** System architecture framework

The two models trained will provide their prediction output to the application layer of the design which is illustrated in Fig. 1.

## *2.3 Application Layer*

In order to achieve the objectives of improving medical adherence and enabling reliable home hospitalization solutions, an application layer is built on top of the developed machine learning models. The application will enable doctors to set up a time schedule to measure the vital signs of the patient depending on the medication schedule. The patient, at home, with the assistance of caregivers if necessary, will key in the vital signs readings to the system at the given time.

Binary predictions will be made based on these inputs, then they will be matched to the schedule of medicine intake. In case the readings showed normal result of the patient vital signs, we will assume that medicine is taken at the right dosage at the right time. Conversely, if the readings are classified as abnormal, they will be presented at the doctor dashboard along with the measured values and the doctor can confirm or deny the prediction. The doctor feedback is then logged and stored in a database for future use as a closed loop system that takes doctors recommendations as input. In case at the next measuring time the results showed abnormal readings again, a red flag will be raised for the doctor to take action to change the dosage of the treatment or the frequency of taking the medicine or they can arrange for appointment with the patient in critical cases.

Multiclass predictions provide an estimate diagnosis of the patient in case the vital signs are not normal. These predictions are then sent to the doctor dashboard when they respond to the patient alert. We do not provide a medical advice based on this; we just provide a predictive information for doctors who can select which prediction is accurate that will be logged into a database and be used again to train the model to provide an improved results in terms of accuracy as more predictions are verified by doctors.

The application of this monitoring process will eventually enable doctors to monitor patients' behavior to treatment while they stay at home. By following their medicine schedule and reporting their vital signs regularly as requested by their doctors, our method can fill in the gap of identifying normal body conditions and reporting any irregular readings of the vital signs where doctors can verify the predictions and request for appointments or change treatment dosage.

## 3 Results and Discussion

### *3.1 Data Preparation*

A well labeled and filtered dataset is prepared which consists of five features and one label. The features included in the dataset are Oxygen Saturation, Blood Pressure (diastolic), Blood pressure (systolic), Heart Rate and Temperature along with a label named condition which identifies whether the readings are considered normal or abnormal.

The prepared dataset consists of the following parameters:

- Blood pressure (diastolic)
- Blood pressure (systolic)
- Heart rate
- Blood oxygen level
- Temperature
- Label.

After filtering patients' readings for abnormal readings to contain only those who have more than 3 readings per day, a total of 1384 rows were extracted. Feature extraction was then performed to obtain the mean, median, min, max and standard deviation. A similar process was done on 102 normal readings that we collected manually with the help of a general physician. Both tables were then combined with the patient ID removed and assigned a label of '0' for normal reading and '1' for abnormal reading.

We are in the process of gathering more data for normal people which will improve system accuracy as the data that we have now is asymmetrical with the larger proportion being the abnormal readings taken from the MIMIC-III dataset and consequently produces less accurate results due to bias caused by the imbalanced data.

### *3.2 Model Training*

Since the prepared dataset consists of 102 rows of normal readings and 1384 rows of abnormal readings, we catered for imbalanced data training. This dataset was

split with the ratio of 50% to 50% for training and testing, respectively. Repeated Stratified K-Fold cross validation was used to split the data with random state set to 1, n_splits set to 4 and n_repeats set to 10 times. One key feature of Repeated Stratified K-Fold cross validation is that it provides balanced weights of training and testing dataset when dealing with imbalanced data as it provides the same proportion of observations with a given categorical value [15]. Another key feature is setting the random state to a specific number which guarantees the same generation of the training and testing sets when adjusting other parameters of the classifiers, so we have a consistent training and testing samples. Splits and repeats setting allows us to train different variations of the same model and recording their accuracy and performance as the training and testing results for large datasets vary according to the sampled data. After splitting the data into training and testing sets, we trained three different models to evaluate their performance.

The first model is k-NN Classifier, which was trained with 'n_neighbors' set to five, 'weights' set to 'uniform', 'algorithm' set to 'auto' and 'metric' set to 'minkowski'. The training produced the following results (Table 1).

The confusion matrix of the obtained results from the k-NN classifier shows good result of predicting abnormal readings but the accuracy of classifying normal readings is low, thus this model is not fit for our goal.

The second model is kernelized SVM. The hyperparameters used were as follows: C set to 1, 'kernel' set to 'rbf, and 'class_weight' set to '{0:0.93, 1:.07}' for imbalanced data. The training produced the following results.

Compared to the previously trained model, kernelized SVM shows a better performance as shown in Table 2. The confusion matrix shows great performance on normal

**Table 1** k-NN classifier test results

| | Normal (predicted) | Abnormal (predicted) | Accuracy (%) |
|---|---|---|---|
| Normal (actual) | 30 | 21 | 58.8 |
| Abnormal (actual) | 13 | 679 | 98.1 |
| Precision | 0.792 | | |
| Recall | 0.748 | | |
| F1 score | 0.975574 | | |

**Table 2** Kernelized support vector classifier test results

| | Normal (predicted) | Abnormal (predicted) | Accuracy (%) |
|---|---|---|---|
| Normal (actual) | 47 | 4 | 92.1 |
| Abnormal (actual) | 89 | 603 | 87.1 |
| Precision | 0.652 | | |
| Recall | 0.870 | | |
| F1 score | 0.8748317 | | |

**Table 3** Linear support vector classifier test results

| | Normal (predicted) | Abnormal (predicted) | Accuracy (%) |
|---|---|---|---|
| Normal (actual) | 44 | 7 | 86.3 |
| Abnormal (actual) | 114 | 578 | 83.5 |
| Precision | 0.540 | | |
| Recall | 0.629 | | |
| F1 score | 0.8371467 | | |

**Table 4** Linear support vector classifier test results

| Model name | Precision | Recall | F1 score |
|---|---|---|---|
| K-NN | 0.792 | 0.748 | 0.975574 |
| SVM SVC | 0.652 | 0.870 | 0.8748317 |
| SVM Linear SVC | 0.540 | 0.629 | 0.8371467 |

and abnormal class in terms of predictions and these result in a better F1 score and recall.

The third model trained is linear SVM configured with 'penalty' set to 12, 'loss' set to 'squared_hinge', 'class_weight' set to 'balanced' and 'intercept_scaling' set to 1. The training process produced the following results.

As showed in Table 3, the linear SVM model showed a moderate performance with slightly decreased accuracy for abnormal class compared to the previous classifier. This was reflected on the F1 score, recall and precision.

Table 4 summarizes the results of these three trained models using two evaluation metrics, Accuracy and F1 score.

By comparing the metrics of the three trained models, as shown in Table 4, SVM SVC is selected based on its accuracy. Further improvements in terms of asymmetrical data training techniques or by adding more data to the collected dataset will be done in the future. The promising results shows the ability to fully integrate the system components and obtain positive results eventually.

## 4 Conclusion and Future Direction

This paper aims to improve medical adherence for patients taking their treatment at home and provides a solution for doctors to keep track of patients' behavior and provide feedback on their condition. We leveraged on supervised learning to train machine learning models with the prepared dataset. The prepared dataset consists of data of vital signs labeled with their medical condition if they are diagnosed with some illness or labeled as normal otherwise. The proposed application that is built on top of these models will provide doctors with a dashboard to monitor patients

medication adherence and provide feedback on the output of the prediction models which can be fed back again to the system input for improved results. We strongly believe that this system will improve medication adherence and provide a method to enable home hospitalization which is on high demand nowadays. The trained model only provide prediction of whether the vital signs are normal or abnormal, however, in order to provide a more helpful estimation, we need to give a score for the degree of which the readings are abnormal. For example, the output could be low, medium, high or normal. In order to further improve the model accuracy, there should be a feedback path for doctors to approve or reject the predictions made by the model which will be used to train and improve another version of the model. We are still collecting data to improve our results and carry out further tests to verify the methodology.

## References


1. Ho PM, Bryson CL, Rumsfeld JS (2009) Medication adherence: its importance in cardiovascular outcomes. Circulation 119(23):3028–3035
2. World Health Organization, Health topics: chronic diseases. [Online]. Available at: http://www.who.int/topics/chronic_diseases/en
3. Span P (2011) A dose of confusion. [Online]. Available at: http://newoldage.blogs.nytimes.com/2011/06/15/a-dose-of-confusion/
4. Polese F, Carrubbo L, Caputo F, Sarno D (2018) Managing healthcare service ecosystems: abstracting a sustainability-based view from hospitalization at home (HaH) practices. Sustainability 10(11):3951
5. Heidenreich PA, Ruggerio CM, Massie BM (1999) Effect of a home monitoring system on hospitalization and resource use for patients with heart failure. Am Heart J 138(4):633–640
6. Tripathi V, Shakeel F (2017) Monitoring health care system using internet of things-an immaculate pairing. In: 2017 International conference on next generation computing and information systems (ICNGCIS). IEEE, pp 153–158
7. Zulkifli, FY, Mustika IW (2018) Development of monitoring and health service information system to support smart health on android platform. In: 2018 4th international conference on nano electronics research and education (ICNERE). IEEE, pp 1–6
8. Hernández C, Aibar J, Seijas N, Puig I, Alonso A, Garcia-Aymerich J, Roca J (2018) Implementation of home hospitalization and early discharge as an integrated care service: a ten years pragmatic assessment. Int J Integr Care 18(2)
9. Lobato SD, Lorenzo FG, Mendieta MG, Alises SM, Arechabala IM, Fernández-Montes CV (2005) Evaluation of a home hospitalization program in patients with exacerbations of chronic obstructive pulmonary disease. Archivos de Bronconeumología ((English Edition)) 41(1):5–10
10. Federman AD, Soones T, DeCherrie LV, Leff B, Siu AL (2018) Association of a bundled hospital-at-home and 30-day postacute transitional care program with clinical outcomes and patient experiences. JAMA Intern Med 178(8):1033–1040
11. Sherif S, Tan WH, Ooi CP, Sherif A, Mansor S (2020) LoRa driven medical adherence system. Bull Electr Eng Inf 9(6):2294–2301
12. Daramola O, Nysaulu P (2019) A digital collaborative framework for improved tuberculosis treatment adherence of patients in rural settings. In: Open innovations (OI), Cape Town, South Africa, pp 297–303. https://doi.org/10.1109/oi.2019.8908213
13. Kumar MP, Nelakuditi UR (2019) IoT and I2C protocol based M-health medication assistive system for elderly people. In: IEEE 16th India Council international conference (INDICON), Rajkot, India, pp 1–4. https://doi.org/10.1109/indicon47234.2019.9030322

14. Nundy S, Patel KK (2020) Hospital-at-home to support COVID-19 surge—time to bring down the walls. (online https://jamanetwork.com/channels/health-forum/fullarticle/2765661)
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, … others (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(Oct):2825–2830

# Finger Vein Presentation Attack Detection Based on Texture Analysis


**Nurul Nabihah Ashari, J. H. Teng, T. S. Ong, and S. M. A. Kalaiarasi**

**Abstract** Biometrics is an effective way to identify and authenticate users based on their personal traits. Among all kinds of hand-based biometrics, finger vein appears to be emerging biometrics that has received a great attention due to its rich information available and ease for implementation. With finger vein system becoming more and more popular, there have been various attempts to comprise the system. Recent studies reveal the vulnerabilities of finger vein system to presentation attack where the sensory device accepts a fake printed finger vein image and gives access as if it were a genuine attempt. In this study, a presentation attack detection method based on hybrid feature spaces of finger vein texture analysis is proposed. Histogram of oriented gradient operator is applied on different channels of grayscale and color feature spaces to obtain texture information of the histogram descriptors. The proposed method includes two implementations of feature space analysis, namely $CHOG_1$ and $CHOG_2$. A well-established publicly available dataset is used to analysis and evaluate the proposed implementations. Experimental results suggest that the combination channels of grayscale and color luminance is able to generate better performance through Support Vector Machine classifier with ACER as low as 0.60% and 0.74% for $CHOG_1$ and $CHOG_2$, respectively. The experiments show that the implementation of $CHOG_1$ performs slightly better than single channel max gradients of $CHOG_2$.

**Keywords** Presentation attack detection · Texture analysis · HOG

## 1 Introduction

Biometric provides an automatic personal recognition of human anatomic or behavioural characteristics. Some common biometrics for user recognition today include fingerprint, iris, face and others. Among different kinds of hand biometrics, finger vein is emerging as one of the popular biometrics chosen as finger vein pattern resists to change over a lifetime and each finger vein pattern is unique and permanent

N. N. Ashari · J. H. Teng · T. S. Ong (✉) · S. M. A. Kalaiarasi
Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia
e-mail: tsong@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_35

[1]. Unfortunately, recent studies revealed that the finger vein system can be susceptible to presentation attack such that printed images can be used to counterfeit finger vein sensor and gain unauthorised access to the system [2]. Therefore, there is a need to ensure finger vein technology can be deployed securely in highly confidential area such as financial, immigration, access control system etc.

Typically, image acquisition is the first step where image of finger vein is acquired by near infra-red (NIR) light. The device consists of putting together a NIR part for the finger placement and a charge-coupled device (CCD) pre-processor camera to capture the image. The vulnerability of presentation attack is possible to happen during the image acquisition step. Tome et al. [2] presented finger vein spoofing starts with enhancing the original image, rescaling them to the original size, printed on high quality paper and presented to fake the sensory device. By looking at presentation attack problem which also known as spoofing attack, this study aims to design a secure finger vein presentation attack detection (PAD) method. The results showed that the attack has a spoofing false accept rate (SFAR) of 76%, which proves that typical finger vein recognition system is vulnerable to spoofing attacks. The work demonstrated even though anti-spoofing method must be put in place to prevent it from happening. Nevertheless, there is still a limited study on the topic.

In this paper, we investigate the extension of histogram of oriented gradients (HOG) operator on different feature spaces of finger vein texture analysis for PAD. The proposed method will undergo pre-processing of segmentation using Watershed transform and feature space conversion. It then followed by Histogram of Oriented Gradient (HOG) as the operator to extract texture information from the combination of grayscale and color feature spaces for classification purpose. We evaluate the proposed method by using a standard public SCUT-FVD presentation attack dataset. The paper is organised with Sect. 2 to present the related works for anti-spoofing attacks and finger vein PAD methods. It then followed by Sect. 3 to detail the proposed solutions. Our experimental analysis will be discussed in Sect. 4 and the paper is concluded in Sect. 5 with the suggestions of possible future work.

## 2 Literature Review

To address image spoofing attacks, many researchers have come up with the use of color space information. From the work of Costa et al. [3], the authors proposed color histogram method to detect fake wine via cork images and replay attack face images. This technique utilises two color spaces of YCrCb and CIE Lab. After acquiring the input image, original RGB color space was converted to these two color spaces. RGB is not a good choice in image processing because the information of red, green, and blue get in the way of separating chrominance and luminance information. After converting the input image into different color spaces, it goes through histogram calculation according to the component of the color space. The histogram is produced from different $YC_rC_b$ and Lab color channels. Next, all of the histograms are concatenated into a feature vector and fed into Extra Trees Classifier

to detect whether the image is a genuine or a spoofing attempt. The idea of this methodology stems from the understanding that different images of the same object have dissimilar visual feature. By observing the color histogram output, there are obvious differences in printed and genuine ones—in both YCrCb and Lab color space. Luminance information of YCrCb and Lab color space is regularly shape in genuine cork image meanwhile it is not the case in the printed attack cork image. On the other hand, the chrominance information of printed image is in regular shape compared to the real live image. Based on the experimental analysis, it was found that the HTER for this method with Replay-Attack database is 0.59% and EER is 0.074%, respectively. In addition, color space analysis has also been widely used in other anti-spoofing face detection such as in [4–6] etc.

In general, finger vein PAD methods can be broadly divided into two different categories, namely, texture based and live detection based. We will discuss the related works of texture-based method since our work will be focus on feature space analysis. One of the promising texture analysis method to detect finger vein presentation attack proposed by Qiu et al. (2018) [7] is through Total Variation Decomposition (TVD) method. According to analysis by Qiu et al., it is concluded that methods that focus around texture perform well as the fake images have vaguely different information that can be the factor or feature to be used in separating the real and forged finger vein images. In this paper, the authors do not perform pre-processing to avoid loss of discriminative information for PAD. Though the image from OHP printed finger vein looks similar with the real image according to human eyes, there are differences in terms of image visual quality, blurriness and noise level. In signal processing, Total Variation (TV) regularisation is a denoising algorithm that provides a solution to reduce noise and make the edges smooth even in low signal-to-noise ratios. Local binary pattern (LBP) is then used as the operator to extract distinctive feature from the decomposed components. After performing TV regularisation and LBP descriptor to the images, histogram of the images can be viewed for visual analysis. Finally, Cascaded Support Vector Machine (SVM) was used to classify the image into a spoof or genuine class. In this research, a self-collected dataset of SCUT FVD and a public dataset of IDIAP FVD were used for experimental evaluation. The best result of ACER as low as 0.00% could be achieved for both SCUT FVD and IDIAP FVD using TVD with LBP Descriptor. Maser et al. [8] investigated the applicability of Photo Response Non-Uniformity (PRNU) to detect finger vein presentation attacks. PRNU fingerprints were first generated from two public finger vein datasets to classify between real or spoofed images, whereby the later contain very low variability and the scattering of light is anticipated to be more uniform as compared to a real image. The lowest ACER based Normalized Cross Correlation (NCC) similarity measure at 4.39% could be obtained for IDIAP spoofing dataset with the implementation of Wiener filter and zero-mean filter post-processing. On the other hand, Singh [9] presented a novel texture method based on the decomposition of normal-map and diffuse-map properties of each finger vein image for PAD purpose. The proposed approach was evaluated by using self-collected dataset with perfect accuracy. Other popular finger vein texture analysis methods for PAD presented in the literature are [10, 11].

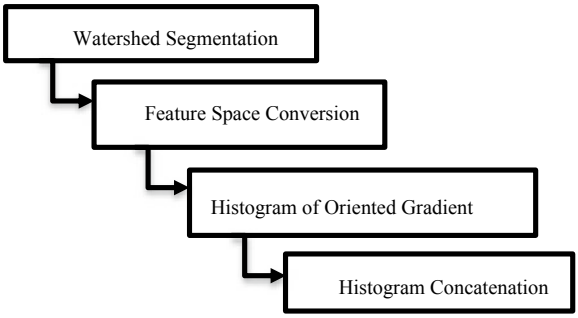



**Fig. 1** Proposed solution

## 3 Proposed Solution

### *3.1 Overview*

In this work, we proposed finger vein PAD based on hybrid texture analysis which includes pre-processing, multi-channel feature space analysis and classification. In this context, the pre-processing step contains watershed algorithm is first applied for image segmentation to obtain the Region of Interest (ROI) of finger vein, It then follows by the conversion of original finger vein image to RGB and transform into the corresponding color space to extract luminance components so that it can be used for subsequent multi-channel analysis, and then Histogram of Oriental Gradient descriptors are calculated from the channels of different feature spaces (grayscale and luminance) to serve as input to be fed into Support Vector Machine (SVM) for classification purpose. The general idea of our proposed methods for detecting presentation attacks is depicted in Fig. 1.

### *3.2 Segmentation*

Watershed segmentation is applied to retrieve the ROI of an image for subsequent analysis. This algorithm makes use of user-defined markers which make it possible to treat pixel values as a local topography (elevation) where the pixel intensity would signify the height. For example, bright areas can be deemed as high area hence labelled as the hill or ridge. For the dark areas on the other hand is deemed as low are and represent trough or basins. It is important to note that Elevation map is computed using a Scharr transform which utilise the optimisation of weighted mean squared angular error.

### 3.3 Feature Space Conversion

In image processing, color space is the color model available for the images. Some famous color space would include RGB (Red, Green, Blue), YUV (Y is the luminance component and UV are chrominance components), YCrCb (Y represents luminance component while Cr and Cb represent chrominance of red and blue respectively) and HSV (hue, saturation, value) [12]. Luminance is a scientific term to describe the brightness of light, while chromium is the color information of an image. According to Li et al. [13], RGB is not the most suitable color space for PAD image analysis due to its inability to separately represent the luminance and chrominance information. As such, we consider YCrCb and YUV color spaces since real and printed images of a subject have different visual features of luminance and chrominance components of the color spaces. YCrCb is chosen because the analysis of the histogram of these color spaces can be an indicator for the liveness of an image [3], and human visual system has different sensitivity towards color and brightness in these color spaces. On the other hand, HSV color space defines an image that can be characterised by its brightness and chromaticity family. Brightness is the measurement of luminous intensity value, while chromaticity is described by hue and saturation. The differences between a real image and a printed image are best described by decoupling the intensity value from color carrying luminance information [14]. In this work, we extract luminance component of the corresponding color spaces and combine with original grayscale texture for subsequent multi-channel feature space analysis.

### 3.4 Histogram of Oriental Gradient (HOG) Feature Space Analysis

Histogram of Oriental Gradient (HOG) is a feature descriptor provides a simplified representation of an image for object detection and image classification. HOG descriptor focuses on the structure and shape of an image object. It can determine the edges of the image through extracting the gradient and orientation of the edges. In our work, locally converted feature space information of an image is first used to calculate gradients on each image pixel. The gradient magnitude of the pixel is added to the corresponding orientation bin. Input pixels are spatially quantized in cells of $n \times n$ pixels, where $n$ is the cell size. In each cell, we compute a 1D histogram of the gradient orientations binned into $b$ bins with tri-linear interpolation on the image. Noted that $b$ and $n$ will be optimized based on empirical evaluation. The final histogram, $H$ is then computed based on block normalised gradient orientation of all the cells of the image. To obtain hybrid histogram descriptors, two implementations, namely $\text{CHOG}_1$ and $\text{CHOG}_2$ are used to represent and analysis different channels of feature spaces for HOG calculation.

Specifically, the implementation is based on two different feature spaces as mentioned in the Sect. 3.3. Let $I$ be a finger vein image represented in the feature

space $C$ such that $C \in \{Grayscale, YCbCr, HSV, YUV\}$ and let $H_{c,}^{i}$ $\{i = 1, \ldots, M\}$ be the histogram extracted from $M$ channel of different feature spaces of $C$. The HOG of the image $I$ can be defined as $\left[H_c = H_c^1, \ldots, H_c^M\right]$, which represents the first implementation where HOG operator is applied to selected feature spaces to obtain final histogram descriptors, $CHOG_1$. In second $CHOG_2$ implementation, multiple channels are transformed into a single channel with maximum channel gradient, which is defined to be the gradient with largest L2 magnitude among the different channels at the pixel. HOG operator is applied to the selected single channel for convenience of histogram calculation.

## 4 Experimental Setup

### *4.1 Datasets and Performance Criteria*

SCUT Finger Vein Presentation Attack Database (SCUT-FVD) collected by BIP Lab, School of Automation Science and Engineering, South China University of Technology (SCUT) [15] will be used for experimental analysis purpose. Note that SCUT-FVD follows the standard protocol to divide the images into three sub-group with reference to International competition of finger vein PAD [11] based on the ratio of sample sizes set as 2:2:6. The images of SCUT-FVD is split into the sub-groups made up of:

(1) Training dataset ("train") contains 1440 total images, half of them are real images, and another half is spoof images. The images are used to train the classifiers used in the study.
(2) Development dataset ("dev") contains 1440 total images, 720 of the images is in real folder, and another 720 are spoof images. Development dataset enables the experiment to estimate decision threshold for SVM. Bayesian Hyper-parameter tuning is used to select a set of optimal hyperparameters with Tree-structured Parzen Estimator (TPE) model as surrogate function to map the hyperparameters to the probability of score of development dataset. The optimal hyperparameters will then be used in test dataset to predict the error rate.
(3) Test dataset ("test") has the largest number of images that is totalled to 4320 images where 2160 of them are real images and another 2160 are spoof images. The test images are used to predict the performance of proposed implementations.

As shown in the Table 1, there are 100 subjects for each real and spoof image of three fingers (index, middle, and ring finger) from both hands, yielding the total of $100 \times 2 \times 3 \times 6$ shoots $= 3600$ images for each subset of real and spoof database, respectively. A subject can only exist in one of the data sets (train, dev or test) to ensure fair comparison purpose. Images are labelled as the following pattern: ID_finger_session_shot_light.bmp. ID is the subject's ID, types of finger of both

**Table 1** SCUT-FVD dataset information

| | Subject | Hands | Fingers | Shot number | Total |
|---|---|---|---|---|---|
| Real | 100 | 2 | 3 (index, middle, ring) | 6 | 3600 |
| Spoof | 100 | 2 | 3 (index, middle, ring) | 6 | 3600 |

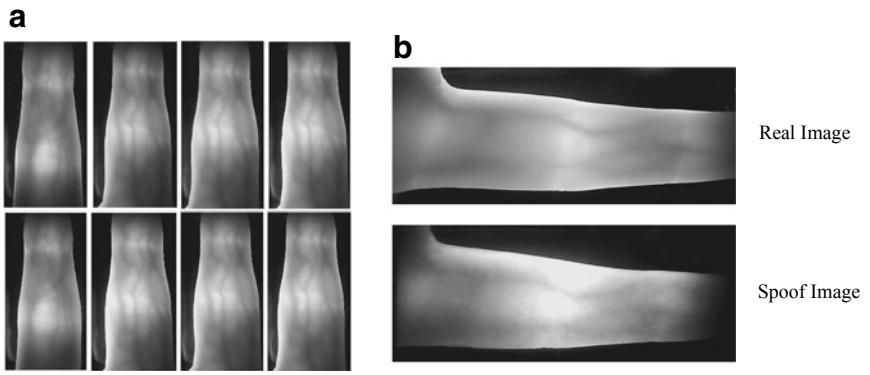



**Fig. 2** **a** Sample images of SCUT FV dataset, **b** real and spoof images of the same finger

hands, the session number (either 0 or 1), shot number, and lastly the intensity of the light for the picture. Figure 2a shows the sample of finger vein images of SCUT FVD while Fig. 2b shows the different types of real and spoof images of the same finger available in the dataset.

ISO/IEC 30107-3 [16] performance criteria is used to evaluate the proposed methods and compare to other different PAD methods. Two evaluation metrics are Bona Fide Presentation Classification Error Rate (BPCER) and Attack Presentation Classification Error Rate (APCER). Specifically, BPCER measures the proportion of bona fide (real) images incorrectly classified as presentation attack images and APCER measures the proportion of attack presentations incorrectly classified as bona fide presentations, while Average-Classification-Error-Rate and (ACER) is calculated as the average error of APCER and BPCER. A low ACER is desired for any PAD methods.

## *4.2 Experimental Results*

Table 2 presents the results of different feature space implementations. The result vindicated that the use of hybrid textures yielded excellent results in finger vein anti-spoofing. From Table 2, it can be seen that the combination of Grayscale and YUV (with luminance component) feature space performs better among the two implementations. In addition, the experiments reveal that $CHOG_1$ provide a slight

**Table 2** Result of the experiments

| Implementation with different feature spaces | Performance criteria (%) | | | |
|---|---|---|---|---|
| | Accuracy | APCER | BPCER | ACER |
| $CHOG_1$-Gray | 98.89 | 1.53 | 0.49 | 1.01 |
| $CHOG_1$-Gray + HSV | 98.84 | 2.08 | 0.23 | 1.16 |
| $CHOG_1$-Gray + YUV | 99.40 | 1.11 | 0.09 | **0.60** |
| $CHOG_1$-Gray + YCbCr | 99.00 | 1.67 | 0.32 | 1.00 |
| $CHOG_2$-Gray | 98.35 | 1.48 | 1.82 | 1.65 |
| $CHOG_2$-Gray + HSV | 99.26 | 1.34 | 0.14 | **0.74** |
| $CHOG_2$-Gray + YUV | 97.06 | 4.68 | 1.20 | 2.94 |
| $CHOG_2$-Gray + YCbCr | 97.75 | 2.92 | 1.57 | 2.25 |

better result than using single-channel max gradients of $CHOG_2$. It is also noticeable that the effect of HOG with different feature spaces could attain the lowest ACER of 0.60 and 0.74% in both $CHOG_1$ (Grayscale + YUV) and $CHOG_2$ (Grayscale + HSV) implementation, respectively. It seems that applying the combination of different feature spaces to be more effective against finger vein presentation attacks. This is because the use of two feature spaces provide enough discriminative information to distinguish the images of spoofed class from real class. On the other hand, Fig. 3 shows Receiver Operating Characteristic (ROC) plot to compare the performance of two implementations with eight feature spaces analysis. The ROC justified the claim of combination of two feature spaces always outweigh single feature space (Grayscale only) with the curve of different two feature spaces always peak at upper left corner on top of single feature space.

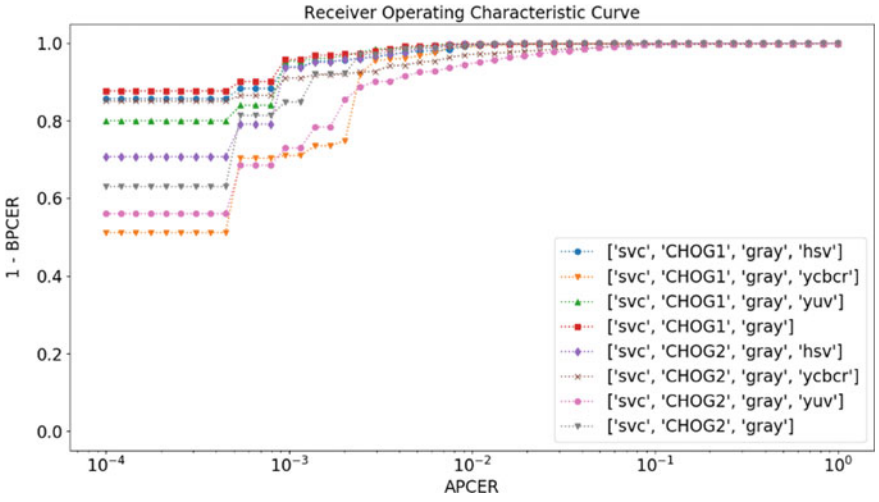



**Fig. 3** ROC plot based on different feature space analysis

**Table 3** Performance comparison of different PAD methods with SCUT-FVD

| Methods | Performance criteria (%) | | |
|---|---|---|---|
| | APCER | BPCER | ACER |
| TV-LBP [7] | 0.00 | 0.00 | 0.00 |
| Proposed $CHOG_1$-Gray + YUV | 1.11 | 0.09 | **0.60** |
| Proposed $CHOG_2$-Gray + HSV | 1.34 | 0.14 | **0.74** |
| DDWT [10] | 2.92 | 0.28 | 1.60 |
| RLBP [11] | 2.18 | 1.67 | 1.93 |
| FSER-DWT [10] | 4.03 | 0.23 | 2.13 |
| HDWT [10] | 9.54 | 1.85 | 5.69 |

Table 3 presents the comparison between our best implementation with other the state-of-the-art methods. As indicated in this table, we observed that TV-LBP is the best PAD method with ACER at 0% while our proposed implementations attain second and third place on the challenging SCUT FVD, which outperform other state of-the art methods. The results reflect that the proposed combination of different feature spaces is capable of attaining the aims of designing a robust finger vein PAD method.

## 5 Conclusion

The work presented multi-channel HOG with two implementations, namely, $CHOG_1$ and $CHOG_2$. The proposed implementation is evaluated by using standard public dataset of SCUT-FVD with promising results, as compared with other state-of-the-art methods. As of the future work, we will extend our works to other histogram descriptors such as multi-resolution HOG and other variants of histogram descriptors for Finger vein PAD problems.

**Acknowledgements** The research is supported by MOHE Fundamental Research Grant Scheme Malaysia (FRGS Grant No: MMUE/190047). Special thanks for BIP-Lab to share SCUT-FVD for experimental analysis purpose.

## References

1. Shaheed K, Liu H, Yang G, Qureshi I, Gou J, Yin Y (2018) A systematic review of finger vein recognition techniques. Information 9(9):213. MDPI, Switzerland
2. Tome P, Vanoni M, Marcel S (2014) On the vulnerability of finger vein recognition to spoofing. In: 2014 international conference of the biometrics special interest group (BIOSIG), Darmstadt, pp 1–10

3. Costa V, Sousa A, Reis A (2018) Image-based object spoofing detection. Lecture Notes in Computer Science, vol 11255. Springer, Heidelberg, pp 189–201
4. Boulkenafet Z, Komulainen J, Hadid A (2015) Face anti-spoofing based on color texture analysis. In: IEEE international conference on image processing (ICIP). Quebec City, Canada, pp. 2636–2640
5. Lu Z, Jiang X, Kot A. (2017) A novel LBP-based color descriptor for face recognition, In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, pp 1857–1861
6. Wen D, Han H, Jain AK (2015) Face spoof detection with image distortion analysis. IEEE Trans Inf Forensics Secur 10(4):746–761
7. Qiu X, Kang W, Tian S, Jia W, Huang Z (2018) Finger vein presentation attack detection using total variation decomposition. IEEE Trans Inf Forensics Secur
8. Maser B, Sollinger D, Uhl A (2019) PRNU-based detection of finger vein presentation attacks. In: 7th international workshop on biometrics and forensics (IWBF), Cancun, Mexico
9. Singh M, Venkatesh S, Raja KB, Ramachandra R, Busch C (2019) Detecting finger-vein presentation attacks using 3D shape & diffuse reflectance decomposition. In: 15th international conference on signal-image technology & Internet-based systems (SITIS), Sorrento, Italy
10. Nguyen DT, Park YH, Shin KY, Kwon SY, Lee HC, Park KR (2013) Fake finger-vein image detection based on fourier and wavelet transforms. Digit Signal Process 23(5):1401–1413
11. Tome P et al (2015) The 1st competition on counter measures to finger vein spoofing attacks. In: Proceedings of the international conference on biometrics (ICB), May 2015, pp 513–518 (2015)
12. Lukac R, Plataniotis KN (2007) Color image processing: methods and applications, Image Processing Series, vol 7. CRC Press, New York
13. Li L, Correia PL, Hadid A (2018) Face recognition under spoofing attacks: countermeasures and research directions. IET Biom 7(1):3–14
14. Kerr DA (2020) Chromaticity and chrominance in color definition. http://www.ma.utexas.edu/users/davis/reu/ch1/signals/chroma.pd. Last accessed 2020/30/7
15. BIP-LAB, GitHub—BIP-Lab/SCUT-SFVD: SCUT-SFVD: A finger vein spoofing/presentation attack database. https://github.com/BIP-Lab/SCUT-SFVD. Last accessed 2020/17/7
16. SC37-Biometrics-Presentation Attack Detection. ISO/IEC Standard FDIS 30107-3. https://christoph-busch.de/files/Busch-PAD-standards-170329.pdf. last accessed 2020/31/7

# Modeling Tourism Using Spatial Analysis Based on Social Media Big Data: A Review


**Zhu Chen, Rayner Alfred, and Oliver Valentine Eboy**

**Abstract** Since an ever-increasing part of the population makes use of social media in their day-to-day lives, social media data has been analyzed in many different disciplines. While there is a great deal of literature on the challenges and difficulties involving specific data analysis methods, there hardly exists research on analyzing the appropriate techniques used to handle different types of data for the purpose of social media analytics. To address this gap, we conducted an extended and structured literature analysis through which we identified challenges addressed and solutions proposed. The literature search revealed that three types of data that were least used for social media analytics that includes Bluetooth, WIFI and mobile roaming data. In contrast, other types of data have received more attention. Based on the results of the literature search, we discuss the most important challenges for researchers and present potential solutions. The findings are used to extend an existing framework on social media analytics. The article provides benefits for researchers and practitioners who wish to collect and analysis social media data.

**Keywords** Tourism · Spatial analysis · Social media · Big data

Z. Chen · O. V. Eboy
Geography Program, Faculty of Humanities, Arts and Heritage, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
e-mail: zhu_chen1990@163.com

O. V. Eboy
e-mail: oliver@ums.edu.my

R. Alfred (✉)
Knowledge Technology Research Unit, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
e-mail: ralfred@ums.edu.my

Z. Chen
School of Resource and Environment, Xingtai University, Xingtai, China



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_36

## 1 Introduction

With the rapid development of social media, the tourism industry has undergone tremendous changes [1]. In the undeveloped stage of the Internet, people mainly navigate in unfamiliar environments through public facilities or tools [2]. Before traveling, they can now use social media to get more information and guidance, and to experience the culture and customs of the destination in advance by following to the travel notes posted by others [3]. Along with the deep comprehension, social media has become the main carrier of the digital media information era, and has received extensive support in society [4]. Because of the good interactivity and timeliness of social media, people are more willing to share their lives on the Internet through social media platforms [5].

As more and more people share their lives online, the large-scale data from structured and unstructured is to form big data, opening up the era of big data [6]. However, there are only three articles on applying big data to tourism research [7–9]. The dimensions of the research field are the starting point for these three studies and focus on a specific type of big data. Regarding different travel issues, different data sources have different data characteristics, and the applied analysis methods are also different. Therefore, according to the application of different types of big data in tourism research, systematic analysis should be carried out from research focus, data types, technical methods, and so on.

The main goal of this paper is to provide a comprehensive overview of the application of big data in tourism research. For the existing review articles, the main contributions of this article are: (1) A comprehensive discussion of the types of big data in tourism research; (2) A comprehensive discussion of the visualization of social media big data; (3) A comprehensive discussion of existing social media big data analysis methods.

## 2 Social Media Data

This section covers a brief overview of types of social media data that can be categorized into Users Generated Content (UGC) Data and Device Generated Content (DGC) Data.

### *2.1 Users Generated Content (UGC) Data*

Users Generated Content data mainly includes two types of data, one is travel text data, such as Blogs, Weibo and Raiders shared by tourists on social platforms; the other is online photo data, such as travel photos shared on Flickr and Instagram. These types of data are used as a primary data to promote tourism research.

### 2.1.1 Text Data

Social media provides a platform for visitors to spread a variety of travel-related information, such as travel reviews and experiences. Visitors can express their satisfaction with travel products on social media platforms and share their travel perspectives and experiences. This has resulted in two types of online text data for travel research, namely review data and blog data. The review data is mainly used to evaluate the satisfaction of tourists with tourism products. The review data mainly comes from the text data of tourists on social platforms such as online reviews of travel websites, hotel booking sites, and blogs. The data volume of review data ranges from hundreds to hundreds of thousands [10–13]. The blog data is primarily used for record travel events and feelings of Tourists. The analysis and research on blog data mainly involves travel recommendation [14] and tourist sentiment analysis [15]. In summary, online text data mainly includes online comment data and blog data, which are commonly used to analyze tourists' satisfaction with hotels, restaurants and scenic spots, as well as travel recommendations and tourist sentiments.

### 2.1.2 Online Photo Data

The photo data posted by users on social platforms contains rich information, such as photo annotations, location and time information. Researchers mine this information to analyze tourists' travel behavior and describe tourists themselves [16–18]. Based on photo data, the discussion of travel recommendations for travel destinations, travel routes, travel durations, travel path and travel time have been well studied [19, 20]. By mining and analyzing the content of tourist photos, researchers can construct images that visualize tourist destinations, and the content of these photos has also become an effective tool for potential tourists to construct images of tourist destinations [21].

## *2.2 Device Generated Content (DGC) Data*

### 2.2.1 Global Positioning System (GPS) Data

Since both global and accurate, GPS data fully demonstrates the feasibility and superiority in tourism research [22]. In early research, the feasibility and practicality of GPS data was the main focus of research [23]. In the following research, tourism behavior mainly focused on spatial behavior and time behavior [24]. However, existing research is more inclined to combine spatial behavior with temporal behavior [25]. With the deepening of research, using GPS data to collect trajectory information of tourists, discover tourists' travel routes and behaviors, and model and define tourist routes to predict tourists' behavior and plans in tourist destinations [26].

### 2.2.2 Mobile Network Roaming Data

Mobile network roaming data is in its infancy, mainly concentrated in the investigation of data applicability and tourism behavior research. Researchers from data collection to system analysis, in addition to studying the applicability of mobile network roaming data in tourism research [27], also on tourism destination management [28], tourist travel time issues, tourist destination loyalty and marketing [29], tourists' spatial distribution and accuracy [30]. However, the acquisition of mobile network roaming data is more difficult, because such data involves user privacy and monitoring issues, which greatly limits the application of mobile network roaming data in tourism research. However, mobile network roaming data can provide a new perspective for tourism research because of its large coverage and rich information.

### 2.2.3 Bluetooth Data

The devices that people carry can be detected by sensors to collect Bluetooth data. From the perspective of location and trajectory, Bluetooth technology can detect the personal behavior of a large number of tourists, providing a new perspective for tourism research. Due to the limited range of Bluetooth data reception, current research is focused on specific area or planned event tourism activities [31, 32].

### 2.2.4 WIFI Data

WIFI, also known as wireless hotspot, is a wireless LAN technology. In terms of tracking travel behavior, WIFI is considered a replacement for Bluetooth technology. Compared with the application of Bluetooth data in tourism research, there are only two studies on the application of WIFI data in tourism research [33, 34]. Similar to Bluetooth data, due to user privacy issues, there are relatively few tourism studies using WIFI data.

## *2.3 Transactional Data*

### 2.3.1 Web Search Data

Visitors use search engines to collect information and leave search trails on the website. These search tracks constitute a class of valuable big data and directly reflect the attention of tourists to tourism projects. Web search data shows great advantages in tourism research, especially in capturing online behaviors and making relevant decisions [35]. Researchers mainly based on Google and Baidu search engine to predict the tourist traffic or hotel visitor number [35–43]. In addition, researchers

use search data to predict economic indicators [44], hotel room demand [45], and travel rate at destination [46].

### 2.3.2 Other Transaction Data

Tourist-related operations such as online booking and purchase of travel products are another part of the transaction data. The hotel website is able to record data such as online bookings and purchases of visitors, which can provide decision support for managers and investors [47]. The occupancy data is used to better help decision makers manage hotels [48]. The consumables, such as electricity, disposables and domestic water, can also be used in tourism research [49]. Visitors browse or access data generated can help the website improve online marketing from content and design [50]. When a tourist uses a consumer card to purchase a travel product, the data is preserved and helps to study the purchase behavior of the visitor and design a customized product [51]. Ticketing data for tourist areas can also support tourism research to improve the management of tourist areas [52]. Analysis using road traffic data can reflect the spatial characteristics of visitors [53]. A small number of tourism researches using travel data also include decision support for tourism management and optimization of online travel products. However, most of this data is in the hands of tourism organizations or governments, and it is difficult to obtain such data, which leads to less travel research using these types of data.

## *2.4 Characteristics of Data*

For online textual data, this data comes from different social media platforms such as TripAdvisor [11, 54–56], Yelp [57], Expedia [12], Ctrip [58], Booking [13], Dianpin and so on. As one of the largest travel social platforms, TripAdvisor is the most widely used in tourism research. Sina Weibo and Twitter are the main blog social platforms and the main source of blog data. Sina Weibo is one of the largest social platforms in China and is also used in tourism research [59]. Twitter data are used to study the temporal and spatial distribution of tourists [60], the relationship between tourist areas and local geographic locations [61], and sentiment analysis [15]. Flickr is one of the largest photo sharing platforms and has become the main source of photo data in tourism research [18, 19]. The GPS data in tourism research is mainly obtained by a professional GPS device [25] or a mobile GPS embedded in a smart phone. GPS equipment can ensure data accuracy compared to other methods [26]. However, the GPS equipment is costly and may produce sample bias. Using GPS-embedded mobile applications to obtain visitors' spatial-temporal behavior data is considered a flexible and low-cost method of data acquisition [62]. Google Trends and Baidu Index are the most popular search data sources currently applied to tourism research. But Google Trend is worldwide, Baidu Index only faces to China [43].

## 3 Visualizing Social Media Data

Visualization can help people better understand the analysis results conveyed by social media data [63]. By the triangular frame structure of social network data [64], social media data can be divided into attributes, location and time. From the perspective of attributes, it can be achieved by graphical variables in Bertin theory [65]. In time dimension, time can be described by graphic variables, snapshots or animation techniques, and adding dimensions to describe [66]. In location dimensions, it is necessary to distinguish geographic or graphical location. The geographic location is usually visualized based on the latitude and longitude of the data released; the graphical location is usually visualized based on the connectivity or other characteristics of the data.

The visualization of social media data usually will contain two or three dimensions. However, a method combining the three dimensions has not yet appeared. Geo-visualization is considered a way to fill this gap, and it provides a new method for theoretical construction and understanding of geographic data [67]. The visualization methods could not capture the depth of social media data in the early stages, therefore more advanced geo-visualization methods are needed to reveal social media data in the multidimensional.

The complex structure of social media data creates several key structures of public participation in social media: identity, sharing, communication, and relationships [68]. Due to different levels of service and complexity, social media services integrate these constructions, resulting in different fields and audiences, and forming a highly heterogeneous environment [69]. But there are three commonalities in different constructions, namely the location of nodes, the links between nodes, and the content shared between nodes.

The geo-visualization of network structure focuses on the underlying network structure of social media data. It is achieved by embedding social network graphs into abstract network space. The graph layout algorithm is usually used to organize the nodes and edges, so that each node corresponds to a set of coordinates in the network space and is organized and drawn.

The social network has a higher vitality due to the changes among users, content and links. The dynamic capture poses a challenge to visualization. The three solutions can be applied to the field of dynamic social networks: visualizing the statistical data of the network over time [70], visualizing a series of network snapshots, and using animation technology to visualize the network. In addition, various indicators of social networks can be visualized using visual variables (color, size).

The visualization of social media content also brings challenges to geo-visualization. One of the common methods for visualizing topic information is tag cloud (word cloud). By combining with geographic information visualization, it is possible to communicate the relationship between geographic and subject space. As word cloud applications increase in the visualization of social media content, geographic layout has also emerged [71]. In addition, it is also applicable to video and sound.

The content of social media has higher vitality, because the topics and keywords in the content will change rapidly. Visualizing the spatio-temporal change can discover the spread and change patterns of the theme, and it can also support the prediction of the theme or spread and the evaluation of the results of the activity. Through the coupling of a dynamic activity, dynamic social media content can be visualized. This dynamic activity displays the contribution of social media location as a function of a given time interval and the ranking of users or content at each time interval.

The relationship between geo-visualization and social media data types is shown in Fig. 1. The difference between visualizing network structure and dynamic network structure is that the latter needs to increase the time dimension. In the UGC data, the geo-visualization of the network structure can be performed through the user's social relationship or the relationship mentioned in the content. But in device data and transaction data, this relationship is hidden. Through the visualization of GPS data, an internal network structure can be formed to analyze and discover tourists' behavior and potential points of interest. Because the device and transaction data lack user-published content, the geo-visualization of social media content is mainly for user generated content data.

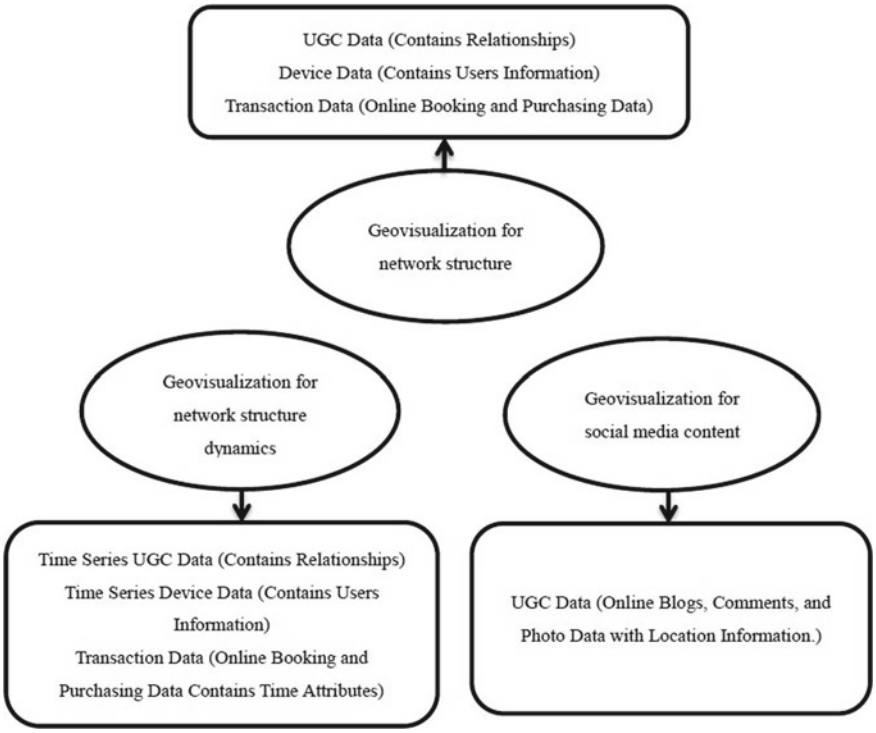



**Fig. 1** Geo-visualization and social media data types. The ellipse represents the type of geo-visualization, and the rectangle represents the type of data that can be used for visualization

## 4 Analyzing Social Media Data

As shown in Table 1, data collecting, and mining are usually used to extract the implicit information in text data. Web crawling technology is used to collect travel-related textual data [10, 14, 72]. In the data mining phase, different processing methods are used according to different research purposes, including data cleaning [10, 12, 13], tokenization [10, 13, 72], word stemming [13] and part-of-speech tagging (POST) [10, 55]. Pattern discovery is another important stage of text data mining including latent Dirichlet allocation (LDA) [10], sentiment analysis [15, 55], clustering and classification [61], statistical analysis [8], text summaries [55] and dependency modeling [54, 58, 73].

The original online photo data is preprocessed by data cleaning and text mining, leaving valuable metadata [19, 74, 75]. After data pre-processing, the cluster analysis of the extracted metadata is mainly carried out from discovery of tourist attractions [19], tourist sources [16] and travel time [20]. Finally, the order of the attractions and the time interval between the attractions are studied to meet the travel route and travel plans [20]. The Markov chain technology is capable of predicting the next destination and obtaining more detailed travel route knowledge based on the location of the visitor [18].

Due to different data formats, the processing of GPS data are divided into direct processing and format conversion of raw GPS data. For the former, techniques such as data cleaning [76], address matching [23, 35] and stop point detection [77] will be performed. After data preprocessing, statistical analysis [78], trajectory clustering [25], motion prediction [26], and frequency pattern mining [77] will be applied to GPS data to study the movement patterns of visitors. For the latter, GPS data is converted into graphic data or vector direction data to study and discover time and space behavior of tourists [77].

The network search data uses keyword selection and prediction factors to achieve tourism prediction in tourism research and the selection method is directly related to the research results [41]. The keyword selection method mainly includes the selection of technical methods [41, 43], region [38] and experience [36]. Most research directly uses raw data as a predictor and introduces network search data of selected keywords into the predictive model [38, 44, 45]. However, in recent years, researchers are more inclined to combine them into one or more composite indices [39, 40, 43].

## 5 Research Limitation

For online text data, the number of dataset samples is relatively small in most studies. In addition, data reliability testing should also be included. Most of the text analysis techniques for a single language are used [10]. In the sentiment analysis, researchers tend to ignore that emotions are affected by verb polarity and degree adverbs [55]. Most of the research lacks understanding and research on the motivation.

**Table 1** Research Method of Applying Social Media Big Data Research

| Data types | Techniques and works |
|---|---|
| Text data | 1. **Web crawling** is used to collect online text data from social platform [10, 14, 72]<br>2. **Data cleaning** is used to detect and remove inaccurate, spelling correction or useless content [10, 12, 13, 80]<br>3. **Tokenization** is used to break text content into meaningful parts [10, 13, 72]<br>4. **Word stemming** is used to identify the root of a word, and mark words [13, 81, 82]<br>5. **Part-of-speech tagging (POST)** is used to tag part of speech for each word [10, 55, 83]<br>6. **Latent Dirichlet allocation (LDA)** is a model for identifying topics in text content [10]<br>7. **Sentiment analysis** is used to identify the tourists' attitudes towards the tourist area [15, 55, 84–86]<br>8. **Clustering and classification** is used to group objects that are similar to each other [61, 87, 88]<br>8. **Statistical analysis** includes descriptive statistics, t-test, correlation matrix, Mann-Whitney U text and correspondence analysis [8]<br>10. **Text summarization** is used to extract key information and generate a summary [55]<br>10. **Dependency modeling** is used to obtain the relationship between tourism factors and text content data [54, 58, 73] |
| Online photo data | 12. **Data cleaning**, **formation** and **text mining** are used to leave useful metadata [19, 74, 75]<br>13. **Cluster analysis**, mainly include centroid-based method, density-based method and connectivity-based method (hierarchical cluster), used to analyses metadata from spatial, temporal and user dimensions [19, 74, 75]<br>14. **Travel routes generation method and Markov chain technique**, used to get travel trajectories [18, 20] |
| GPS data | 15. **Data cleaning** is used to remove noise points in order to improve data accuracy [76]<br>16. **Address matching** is used to project GPS data onto the road network [23, 35]<br>17. **Stop point detection** is used to discover where tourists stay for a while [77, 79]<br>18. **Statistical analysis** is used to capture tourists route based on GPS data [78]<br>19. **Trajectory clustering** is used to aggregate similar paths into a cluster [25]<br>20. **Frequency pattern mining** is used to discover generalized sequences of time-sequential event sets [77]<br>21. **Motion prediction** is used to predict tourists' next destination based on GPS data [26] |
| Web search data | 22. **Three keywords selection methods** mainly include experiential, territorial and technological method [36, 38, 41, 43]<br>23. **Predictor introduction**, include using the raw web search data and index construction such as combining them into a few composite indexes [38–40, 43–45] |

In photo data, there is a lack of relevant research on the time dimension, and the data source is limited to mainstream photo-sharing platforms. In addition to Cluster analysis and ranking analysis, other big data mining techniques should also be introduced into current research. Among the researches, there is a lack of attention to the valuable information contained in the photo data itself.

In GPS data, a few studies focus on GPS data. There is a lack of interdisciplinary research methods to study and analyze GPS data. Forecasting future tourism activities with periodic tourism behavior deserves more attention [79].

The research on tourism market factors is relatively rare by using network search. Most of the research relies on experience or regional methods to select keywords, but both methods have defects. When there are many keywords, more indexes should be constructed to avoid collinearity in the regression.

In general, the current research usually involves only a single type of data and mostly from a single space, lack of grasp of the overall space, especially in different spatial scales.

## 6 Conclusion

The various types used in tourism research mainly come from three aspects. The data generated by users is the main data type applied in tourism research. The equipment data is relatively small and tourism research using operational data accounts for the least, mainly because of data acquisition difficulties. It is not difficult to find that the focus of tourism research on applying big data depends to a large extent on data characteristics. However, different types of big data are used to study the same problem, and each type of big data provides a unique analytical perspective. This is due to differences in data characteristics that lead to their different performance in tourism research.

The high degree of interconnection of social media data makes it necessary to build networks from multiple dimensions. Therefore, analyzing and visualizing these contents can discover valuable knowledge from the various attributes of the data. The visualization of social media resources requires the visualization of abstract concepts. This provides a unique opportunity to combine qualitative and quantitative analysis in geography to study the human social system. Geographical visualization is an attempt to study the complex and multidimensional social system of human beings and will continue to develop with the improvement of analytical capabilities.

The use of big data in travel research faces challenges such as privacy, data quality and data costs. Collaboration between research institutions and tourism agencies may be a way to address these challenges, not only reducing data cost and data availability, but also actually addressing tourism-related issues. The solution to privacy issues can be achieved by signing a confidentiality agreement or culling sensitive information in the data.

## References


1. Hua LY, Ramayah T, Ping TA, Jun-Hwa C (2017) Social media as a tool to help select tourism destinations: the case of Malaysia. Inf Syst Manag 34(3):265–279
2. Berno T, Ward C (2005) Innocence abroad: a pocket guide to psychological research on tourism. Am Psychol 60(6):593–600
3. Kim J, Fesenmaier DR (2017) Sharing tourism experiences: the posttrip experience. J Travel Res 56(1):28–40
4. DeAndrea DC, Ellison NB, LaRose R, Steinfield C, Fiore A (2012) Serious social media: on the use of social media for improving students' adjustment to college. The Intern High Educ 15(1):15–23
5. Liu DJ, Hu J, Cheng SW, Chen JZ, Zhang Q (2015) Spatial distribution pattern and influencing factors of China's tourism Weibo——taking Sina Travel Weibo as an example. Scientia Geographica Sinica 35(06):717–724
6. Kambatla K, Kollias G, Kumar V, Grama A (2014) Trends in big data analytics. J Parall Distrib Comput 74(7):2561–2573
7. Rashidi TH, Abbasi A, Maghrebi M, Hasan S, Waller TS (2017) Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. Transp Res Part C: Emerg Technol 75:197–211
8. Schuckert M, Liu X, Law R (2015) Hospitality and tourism online reviews: recent trends and future directions. J Travel Tour Market 32(5):608–621
9. Shoval N, Ahas R (2016) The use of tracking technologies in tourism research: the first decade. Tour Geogr 18(5):587–606
10. Guo Y, Barnes SJ, Jia Q (2017) Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. Tour Manag 59:467–483
11. Liu Y, Teichert T, Rossi M, Li H, Hu F (2017) Big data for big insights: investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. Tour Manag 59:554–563
12. Xiang Z, Schwartz Z, Gerdes JH Jr, Uysal M (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? Int J Hosp Manag 44:120–130
13. Xu X, Li Y (2016) The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: a text mining approach. Int J Hosp Manag 55:57–69
14. Yuan H, Xu H, Qian Y, Li Y (2016) Make your travel smarter: summarizing urban tourism information from massive blog data. Int J Inf Manage 36(6):1306–1319
15. Philander K, Zhong Y (2016) Twitter sentiment analysis: capturing sentiment from integrated resort tweets. Int J Hosp Manag 55(2016):16–24
16. Da Rugna J, Chareyron G, Branchet B (2012) Tourist behavior analysis through geotagged photographies: a method to identify the country of origin. In: 2012 IEEE 13th international symposium on computational intelligence and informatics (CINTI). IEEE, pp 347–351
17. Lu D, Wu R, Sang J (2017) Overlapped user-based comparative study on photo-sharing websites. Inf Sci 376:54–70
18. Vu HQ, Li G, Law R, Ye BH (2015) Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. Tour Manag 46:222–232
19. Lee I, Cai G, Lee K (2014) Exploration of geo-tagged photos through data mining approaches. Expert Syst Appl 41(2):397–405
20. Lu X, Wang C, Yang JM, Pang Y, Zhang L (2010) Photo2trip: generating travel routes from geo-tagged photos for trip planning. In: Proceedings of the 18th ACM international conference on multimedia. ACM, pp 143–152
21. Deng N, Li XR (2018) Feeling a destination through the "right" photos: a machine learning model for DMOs' photo selection. Tour Manag 65:267–278
22. Bauder M, Freytag T (2015) Visitor mobility in the city and the effects of travel preparation. Tour Geogr 17(5):682–700
23. East D, Osborne P, Kemp S, Woodfine T (2017) Combining GPS & survey data improves understanding of visitor behaviour. Tour Manag 61:307–320

24. Shoval N, McKercher B, Birenboim A, Ng E (2015) The application of a sequence alignment method to the creation of typologies of tourist activity in time and space. Environ Plann B: Plann Des 42(1):76–94
25. Zakrisson I, Zillinger M (2012) Emotions in motion: tourist experiences in time and space. Curr Issues Tour 15(6):505–523
26. Zheng W, Huang X, Li Y (2017) Understanding the tourist mobility using GPS: where is the next place? Tour Manag 59:267–280
27. Ahas R, Aasa A, Roose A, Mark Ü, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: an Estonian case study. Tour Manag 29(3):469–486
28. Raun J, Ahas R, Tiru M (2016) Measuring tourism destinations using mobile tracking data. Tour Manag 57:202–212
29. Nilbe K, Ahas R, Silm S (2014) Evaluating the travel distances of events visitors and regular visitors using mobile positioning data: the case of Estonia. J Urban Technol 21(2):91–107
30. Ahas R, Aasa A, Mark Ü, Pae T, Kull A (2007) Seasonal tourism spaces in Estonia: case study with mobile positioning data. Tour Manag 28(3):898–910
31. Versichele M, De Groote L, Bouuaert MC, Neutens T, Moerman I, Van de Weghe N (2014) Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: a case study of Ghent, Belgium. Tour Manag 44:67–81
32. Yoshimura Y, Sobolevsky S, Ratti C, Girardin F, Carrascal JP, Blat J, Sinatra R (2014) An analysis of visitors' behavior in the Louvre Museum: a study using Bluetooth data. Environ Plann B: Plann Des 41(6):1113–1131
33. Bonné B, Barzan A, Quax P, Lamotte W (2013) WiFiPi: involuntary tracking of visitors at mass events. In: 2013 IEEE 14th international symposium on "A world of wireless, mobile and multimedia networks" (WoWMoM). IEEE, pp 1–6
34. Chilipirea C, Petre AC, Dobre C, van Steen M (2016) Presumably simple: monitoring crowds using WiFi. In: 2016 17th IEEE International Conference on Mobile Data Management (MDM) vol 1. IEEE, pp 220–225
35. Li X, Wu Q, Peng G, Lv B (2016) Tourism forecasting by search engine data with noise-processing. Afr J Bus Manage 10(6):114–130
36. Bangwayo-Skeete PF, Skeete RW (2015) Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tour Manag 46:454–464
37. Gunter U, Önder I (2016) Forecasting city arrivals with Google analytics. Ann Tour Res 61:199–212
38. Huang X, Zhang L, Ding Y (2017) The Baidu index: uses in predicting tourism flows–a case study of the Forbidden City. Tour Manag 58:301–306
39. Li X, Pan B, Law R, Huang X (2017) Forecasting tourism demand with composite search index. Tour Manag 59:57–66
40. Park S, Lee J, Song W (2017) Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. J Travel Tour Market 34(3):357–368
41. Peng G, Liu Y, Wang J, Gu J (2017) Analysis of the prediction capability of web search data based on the HE-TDC method-prediction of the volume of daily tourism visitors. J Syst Sci Syst Eng 26(2):163–182
42. Rivera R (2016) A dynamic linear model to forecast hotel registrations in Puerto Rico using Google trends data. Tour Manag 57:12–20
43. Yang X, Pan B, Evans JA, Lv B (2015) Forecasting Chinese tourist volume with search engine data. Tour Manag 46:386–397
44. Choi H, Varian H (2012) Predicting the present with Google trends. Econ Rec 88:2–9
45. Pan B, Chenguang Wu D, Song H (2012) Forecasting hotel room demand using search engine data. J Hosp Tour Technol 3(3):196–210
46. Gawlik E, Kabaria H, Kaur S (2011) Predicting tourism trends with Google insights. Accessed December 1, 2012
47. Saito T, Takahashi A, Tsuda H (2016) Optimal room charge and expected sales under discrete choice models with limited capacity. Int J Hosp Manag 57:116–131

48. Falk M (2010) A dynamic panel data analysis of snow depth and winter tourism. Tour Manag 31(6):912–924
49. Kahn ME, Liu P (2016) Utilizing "Big Data" to improve the hotel sector's energy efficiency: lessons from recent economics research. Cornell Hosp Quart 57(2):202–210
50. Plaza B (2011) Google analytics for measuring website performance. Tour Manag 32(3):477–481
51. Sobolevsky S, Sitko I, Des Combes RT, Hawelka B, Arias JM, Ratti C (2014) Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain. In: 2014 IEEE international congress on big data. IEEE, pp 136–143
52. Shih C, Nicholls S, Holecek DF (2009) Impact of weather on downhill ski lift ticket sales. J Travel Res 47(3):359–372
53. Huang Z, Cao F, Jin C, Yu Z, Huang R (2017) Carbon emission flow from self-driving tours and its spatial relationship with scenic spots–a traffic-related big data method. J Clean Prod 142:946–955
54. Fang B, Ye Q, Kucukusta D, Law R (2016) Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. Tour Manag 52:498–506
55. Hu YH, Chen YL, Chou HL (2017) Opinion mining from online hotel reviews–a text summarization approach. Inf Process Manage 53(2):436–449
56. Ma J, Luo S, Yao J, Cheng S, Chen X (2016) Efficient opinion summarization on comments with online-LDA. Int J Comput Commun Control 11(3):414–427
57. Park S, Nicolau JL (2015) Asymmetric effects of online consumer reviews. Ann Tour Res 50:67–83
58. Ye Q, Law R, Gu B (2009) The impact of online user reviews on hotel room sales. Int J Hosp Manag 28(1):180–182
59. Cheng M, Edwards D (2015) Social media in tourism: a visual analytic approach. Curr Issues Tour 18(11):1080–1087
60. Chua A, Servillo L, Marcheggiani E, Moere AV (2016) Mapping Cilento: using geotagged social media data to characterize tourist flows in southern Italy. Tour Manag 57:295–310
61. Bordogna G, Frigerio L, Cuzzocrea A, Psaila G (2016) Clustering geo-tagged tweets for advanced big data analytics. In: 2016 IEEE international congress on Big Data (BigData congress). IEEE, pp 42–51
62. Brovelli MA, Minghini M, Zamboni G (2016) Public participation in GIS via mobile applications. ISPRS J Photogramm Remote Sens 114:306–315
63. Freeman L (2004) The development of social network analysis. A Study in the Sociology of Science, 1, 687
64. Peuquet DJ (1994) It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. Ann Assoc Am Geogr 84(3):441–461
65. Bertin J (1983) Semiology of graphics: diagrams. Networks, Maps 10(00690805.1987), 10438353
66. Gaertler M, Wagner D (2005) A hybrid model for drawing dynamic and evolving graphs. In: International symposium on graph drawing. Springer, Berlin, Heidelberg, pp 189–200
67. MacEachren AM, Kraak MJ (2001) Research challenges in geovisualization. Cartogr Geogr Inf Sci 28(1):3–12
68. Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS (2011) Social media? Get serious! Understanding the functional building blocks of social media. Bus Horiz 54(3):241–251
69. Hanna R, Rohm A, Crittenden VL (2011) We're all connected: the power of the social media ecosystem. Bus Horiz 54(3):265–273
70. Ahn JW, Taieb-Maimon M, Sopan A, Plaisant C, Shneiderman B (2011) Temporal visualization of social network dynamics: prototypes for nation of neighbors. In: International conference on social computing, behavioral-cultural modeling, and prediction. Springer, Berlin, Heidelberg, pp 309–316

71. De Chiara D, Del Fatto V, Sebillo M, Tortora G, Vitiello G (2012) Tag@ map: a web-based application for visually analyzing geographic information through georeferenced tag clouds. In: International symposium on web and wireless geographical information systems. Springer, Berlin, Heidelberg, pp 72–81
72. Xiang Z, Du Q, Ma Y, Fan W (2017) A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism. Tour Manag 58:51–65
73. Zhang Y, Cole ST (2016) Dimensions of lodging guest satisfaction among guests with mobility challenges: a mixed-method analysis of web-based texts. Tour Manag 53:13–27
74. Miah SJ, Vu HQ, Gammack J, McGrath M (2017) A big data analytics method for tourist behaviour analysis. Inf Manag 54(6):771–785
75. Oender I (2017) Classifying multi-destination trips in Austria with big data. Tour Manag Perspect 21:54–58
76. Birenboim A, Reinau KH, Shoval N, Harder H (2015) High-resolution measurement and analysis of visitor experiences in time and space: the case of Aalborg zoo in Denmark. The Prof Geogr 67(4):620–629
77. Orellana D, Bregt AK, Ligtenberg A, Wachowicz M (2012) Exploring visitor movement patterns in natural recreational areas. Tour Manag 33(3):672–682
78. Shoval N, McKercher B, Ng E, Birenboim A (2011) Hotel location and tourist activity in cities. Ann Tour Res 38(4):1594–1612
79. Zheng Y (2015) Trajectory data mining: an overview. ACM Trans Intell Syst Technol (TIST) 6(3):29
80. Basri SB, Alfred R, On CK (2012) Automatic spell checker for Malay blog. In: 2012 IEEE international conference on control system, computing and engineering, Penang, pp 506–510. https://doi.org/10.1109/iccsce.2012.6487198
81. Alfred R, Leong LC, On CK, Anthony P (2014) A literature review and discussion of Malay rule—based Affix elimination algorithms. In: Uden L, Wang L, Corchado Rodríguez J, Yang HC, Ting IH (eds) The 8th international conference on knowledge management in organizations. Springer proceedings in complexity. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7287-8_23
82. Leong LC, Basri S, Alfred R (2012) Enhancing Malay stemming algorithm with background knowledge. In: Anthony P, Ishizuka M, Lukose D (eds) PRICAI 2012: trends in artificial intelligence. PRICAI 2012. Lecture Notes in Computer Science, vol 7458. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32695-0_68
83. Alfred R, Mujat A, Obit JH (2013) A ruled-based part of speech (RPOS) tagger for Malay text articles. In: Selamat A, Nguyen NT, Haron H (eds) Intelligent information and database systems. ACIIDS 2013. Lecture Notes in Computer Science, vol 7803. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36543-0_6
84. Wang D, Alfred R (2020) A review on sentiment analysis model for Chinese Weibo text. In: 2020 3rd international conference on advanced electronic materials, computers and software engineering (AEMCSE), Shenzhen, China, pp 456–463. https://doi.org/10.1109/aemcse50948.2020.00105
85. Hung P, Lai, Rayner, Alfred (2019) An optimized multi-layer ensemble framework for sentiment analysis. In: 2019 1st international conference on artificial intelligence and data sciences (AiDAS), Ipoh, Perak, Malaysia, pp 158–163. https://doi.org/10.1109/aidas47888.2019.8970949
86. Alfred R, Teoh RW (2019) Improving topical social media sentiment analysis by correcting unknown words automatically. In: Yap B, Mohamed A, Berry M (eds) Soft computing in data science. SCDS 2018. Communications in computer and information science, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-3441-2_23
87. Marine-Roig E, Clavé SA (2015) Tourism analytics with massive user-generated content: a case study of Barcelona. J Destin Market Manag 4(3):162–172

88. Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F (2017) Natural language processing based features for sarcasm detection: an investigation using bilingual social media texts. In: 2017 8th international conference on information technology (ICIT), Amman, pp 703–709. https://doi.org/10.1109/icitech.2017.8079931

# Analysis of Heart Rate Variability Using Wearable Device


**Rosmina Jaafar and Onn Chung Xian**

**Abstract** Real-life stressors such as work pressure and examination exist in daily life and will affect heart rate (HR) and heart rate variability (HRV) of an individual. The objective of this study is to evaluate the effect of stress on HRV values in healthy human subjects. Wearable device equipped with photoplethysmography (PPG) sensor is worn by all subjects to record the HR data in two situations, which are during rest situation and stress situation for a period of 10 min in each situation. The recorded HR data were then analysed using MATLAB and Kubios HRV software to get the HRV values. Results obtained show that all subjects have higher HRV values during rest situation and those values drop drastically when subjects were exposed to stress simulation. HRV calculated from the root mean square of successive differences (RMSSD) values are more stable and consistent in determining the HRV values compared to standard deviation of the nearest neighbor intervals (SDNN) values. Besides, all subjects show changes in interbeat interval (IBI) from high fluctuations to low fluctuations when subjects are in stress situation. Low fluctuations in IBI changes will result in lower HRV values indicating the presence of stress components. In conclusion, this study provides evidence that one will have lower HRV values during stress situation compared to rest situation.

**Keywords** Heart rate variability · Stress · Photoplethysmography

## 1 Introduction

Wearable sensors have grown significantly in these days and play an important role in human life especially in health monitoring application [1]. Nowadays, wearable sensors not only measure simple parameter such as number of steps taken in a day,

R. Jaafar (✉) · O. Chung Xian
Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia
e-mail: rosmina@ukm.edu.my

O. Chung Xian
e-mail: chungxian19@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_37

but also others physiological data such as heart rate (HR) and heart rate variability (HRV). The normal HR for adults is 60–100 beats per minute (bpm) [2]. Individual will have different HR depending on their body fitness level. Typically, a healthy person will have a lower HR and vice versa.

The two most common ways in measuring HR are by using electrocardiography (ECG) and photoplethysmography (PPG) sensors. Over the years, ECG has been the dominant method in heart rate monitoring to identify abnormality in cardiovascular system. ECG records the electrical activities of the heart and it shows the variation in amplitude of the ECG signals versus time. Even though the ECG technique has been improved progressively for the last few decades, its usage in terms of flexibility and portability still does not meet the expectation of consumers. For example, few electrodes must be placed on the chest area in order for the ECG to function effectively and this procedure greatly reduced the mobility of users.

Thus, PPG is developed as an alternative method in heart rate monitoring [3]. Using comprehensive signal analysis, research has shown that PPG signal has good potential to replace ECG recording in HRV signal extraction [4]. PPG uses infrared light to measure volumetric variation in blood flow [5]. This measurement provides important information on the cardiovascular system. The popularity of using PPG technology as alternative technique in heart rate monitoring has been increased lately, mainly due to easy operation, high flexibility and portability, as well as its availability at an affordable price. Thus, the PPG method is chosen to be used in this study.

While HR focuses on average heartbeat per minute, HRV consists of variation in differences of time between consecutive heartbeats which is also known as the variation in interbeat interval (IBI) [6]. A healthy heartbeat is complex and always changing which enables the cardiovascular system to quickly adapt itself with the physical and psychological challenges until homeostasis is achieved [7]. Typically, a low HRV value shows that the body is under stress while a high HRV value normally indicates that the body has strong ability in handling stress [8].

Mental stress has been a serious problem in modern society. Thus, the ability to monitor stress level can help in overcoming this issue. Currently, the use of HRV to monitor stress is still not developed thoroughly and this condition causes many people not aware of the importance in monitoring stress level [9]. Nowadays, HRV is normally used to identify cardiovascular disease such as heart attack, heart failure and stroke.

Therefore, this study aims to investigate the effect of stress level on the HRV value with the use of wearable sensors to record the HR data. HRV can be presented in time domain, frequency domain, and non-linear methods [10]. This study only focuses on time domain method whereby the values of root mean square of successive differences (RMSSD) and standard deviation of nearest neighbor intervals (SDNN) will be investigated while analyzing the HRV.

# 2 Methodology

## *2.1 Participants*

A random sample of four participants which comprises of three males and one female subject participated in this study. Participants are students at the Faculty of Engineering and Built Environment in Electrical and Electronics Engineering course at the National University of Malaysia (UKM).

Verbal explanation of the study procedure was provided to all participants. All participants were allowed to adapt with the stress simulation task whereby a series of mental arithmetic challenges need to be solved within a time limit. The HR data collection started after the participant had understood all the procedures and has adapted well with the stress simulation task.

## *2.2 Experiment Design and Procedure*

The procedure starts with the HR data collection using photoplethysmography sensors where data were collected before the stress simulation where participants were in rest condition and during stress simulation which indicated the presence of stress element. A suitable time interval in collecting the data was selected so that the analysis can be done accurately. After the data collection has completed, MATLAB and Kubios HRV software were used to analyse the data to obtain the HRV values using statistical method. Lastly, comparison of results between the two softwares were carried out to determine whether their findings comply with each other.

## *2.3 Photoplethysmography (PPG) Sensors*

Huawei Honor Band5 that is equipped with PPG sensor as shown in Fig. 1a (left) was used in collecting HR data. This wearable device uses a LED that shines light into the body tissues and records the amount of light that is reflected back onto the photodiode as shown in Fig. 1b [11].

PPG sensor can operate in two different modes, which are the transmission mode and the reflectance mode respectively. Each mode has its own advantages and disadvantages, but both provide non-invasive measurement to the users.

In the transmission mode, the light source and the photodiode are separated by body tissues whereas in the reflectance mode, the light source and the photodiode are located at the same side with the tissues to measure the reflected light. In this study, the Honor Band5 is using reflectance mode as shown in Fig. 1a (right) to record the HR data.

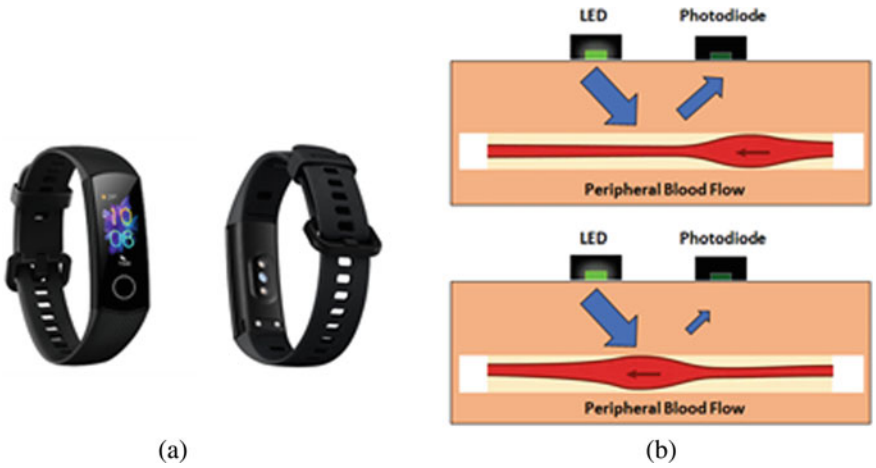



**Fig. 1** **a** Honor band 5 in front view (left), back view (right) and **b** reflectance PPG sensor

## *2.4 Stress Simulation*

Before the stress simulation was conducted, participants were required to be in the sitting position for 10 min to record the HR data in rest condition. In order to get the HR data in stress condition, participants were required to go through a test known as Montreal Imaging Stress Task (MIST) for a time period of 10 min.

MIST consists of a series of computerized mental arithmetic challenges with a time limit for each task [12]. Participants were required to answer a series of mathematical question where the solutions to the questions were integers, between 0 and 9.

Participants will select a number on the rotary dial as shown in Fig. 2 by pressing the left or right mouse buttons. Participants were required to press the middle mouse

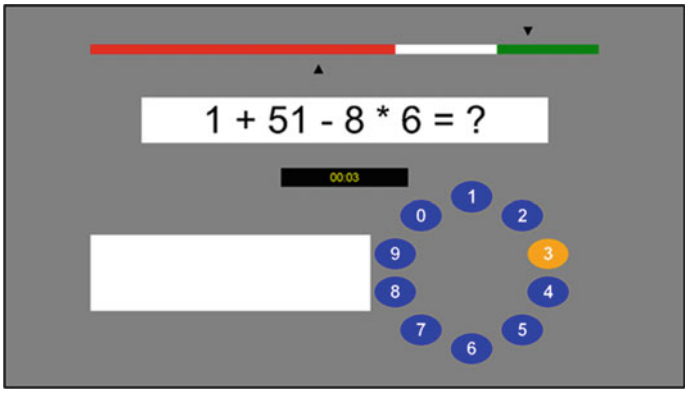



**Fig. 2** User interface in MIST

button to submit the chosen answer on the rotary dial as the response to the arithmetic task. This response was then compared with the correct answer for the task, and the appropriate feedback ("correct" or "incorrect") was shown in the feedback field of the computer screen. If no response was submitted within the time limit, the response "timeout" was displayed.

### *2.5 Statistical Method*

The variations in the HR can be evaluated using few methods. The most common method in calculating the HRV value is by using time or frequency. In this context, the time domain method means the quantity of heartbeat that is formed in a specific period of time, whereas the frequency domain method calculates the quantity of heartbeat with low and high frequencies that have been formed.

In this study, time domain was used in determining the HRV values calculated by. MATLAB programming using the appropriate formula for the RMSSD and SDNN evaluations. The MATLAB coding read the HR data saved in Microsoft Excel downloaded from the PPG sensor in Honor Band5. The HRV data can also be evaluated using Kubios HRV software for comparison [13].

The time domain method can be further divided into statistical method and geometrical method [14]. After taking few aspects into consideration, the statistical method is chosen as the method for evaluations of the HRV. In the statistical method, the most common way that is used in analyzing the HRV is the RMSSD which is easy to calculate and provides an accurate measurement for HRV analysis as it reflects the parasympathetic activities. Besides, another method that is typically used to calculate HRV is the analysis of SDNN which can be calculated using the standard deviation values for all inter-beat intervals [15].

## 3 Results

### *3.1 RMSSD Values*

Equation (1) has been used to calculate the RMSSD value in MATLAB whereas Kubios uses its own algorithm in calculating the RMSSD value.

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2}{N-1}} \tag{1}$$

where

RR indicates the values of peak to peak or the inter-beat interval

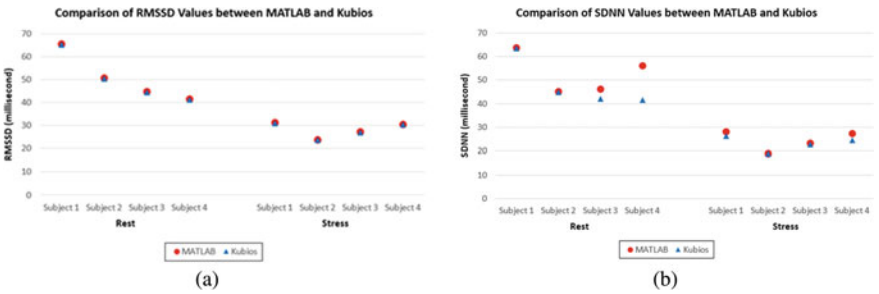



**Fig. 3** HRV using MATLAB and Kubios showing **a** RMSSD and **b** SDNN values

$RR_i$ is the instantaneous inter-beat interval
$RR_{i+1}$ is the next instantaneous inter-beat interval
N is the total number of RR data

Figure 3 shows that the RMSSD values calculated by MATLAB and Kubios software are the same and no obvious differences are observed.

## 3.2 SDNN Values

Equation (2) has been used to calculate the SDNN value in MATLAB whereas Kubios uses its own algorithm in calculating the SDNN value.

$$SDNN = \sqrt{\frac{\sum_{i=1}^{N}\left(RR_i - \overline{RR}\right)^2}{N}} \tag{2}$$

where $\overline{RR}$ indicates the average value of peak to peak or the inter-beat interval.

Figure 3b shows the SDNN values calculated by MATLAB and Kubios software. The SDNN values from Kubios are not consistent and some participants show an obvious difference of SDNN between the two softwares. Refer to Fig. 4b, there are

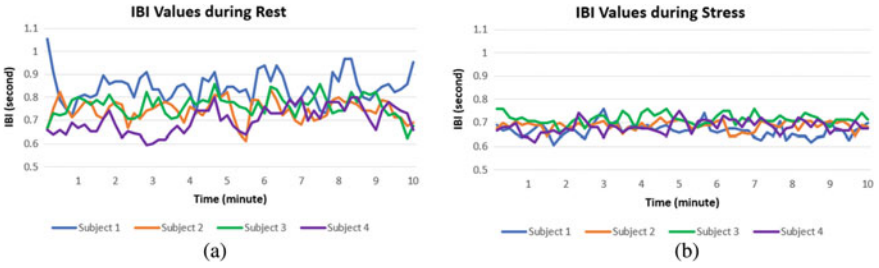



**Fig. 4** **a** IBI values during rest, **b** IBI values during stress

obvious difference in SDNN values for few subjects when the participants are in rest and stress conditions. During rest condition, all participants show a SDNN value of more than 40 ms. On the other hand, all participants show a lower SDNN value which is less than 28 ms during stress condition.

### 3.3 IBI Values

The heart rate (HR) data collected using Honor Band5 produced IBI values that can be calculated using Eq. (3).

$$IBI = \frac{60}{HR} \tag{3}$$

The IBI for all participants in rest situation as shown in Fig. 4a demonstrates high fluctuations in the IBI values. On the other hand, the IBI for all participants show low fluctuations in the IBI values when they are exposed to stress simulation as shown in Fig. 4b. High fluctuations in the IBI values will lead to higher RMSSD and SDNN values and vice versa. The IBI fluctuations is directly proportional to the HRV where high fluctuations during rest indicates high HRV and low fluctuations during stress indicates low HRV.

A box-plot can also be used to show the variation in IBI values. The box-plot shows the distribution of data set based on five number summaries; minimum, first quartile (Q1), median, third quartile (Q3) and maximum. The difference between Q3 and Q1 shows the interquartile range (IQR). Big values of IQR indicates a high variation in a set of data and vice versa. Figure 5 shows all participants data have bigger values in IQR during rest condition compared to stress condition. This indicates that subjects have high variation in the IBI values during rest that will lead to higher HRV values.

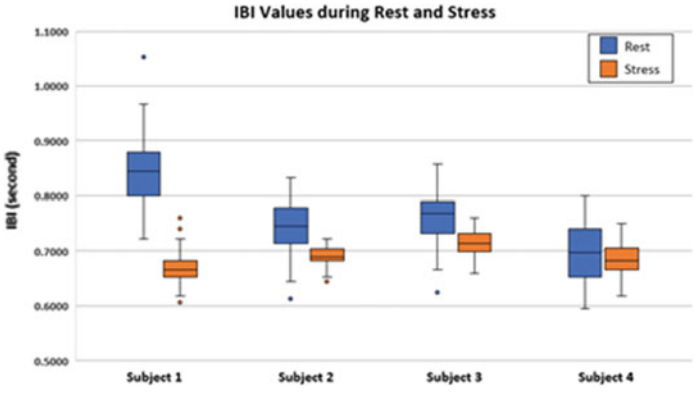



**Fig. 5** Box-plot for IBI values for all subjects

## 4 Discussion

In this study, PPG technique has been used where all participants are required to wear Huawei Honor Band 5 during the HR data collection. HR data is recorded in two conditions, which are when participants are in rest condition and when they are exposed to stress simulation. This study only focuses on two HRV parameters in time domain which are the RMSSD and the SDNN.

The recorded HR data is analyzed using MATLAB and Kubios software in order to get the RMSSD and SDNN values. Results show that the RMSSD values obtained from both software are very similar and consistent for all participants whereas the SDNN values obtained are not consistent and some participants show obvious difference between the two software. This shows that the RMSSD values are more stable and consistent in determining HRV values compared to the SDNN values. However, both software recorded a higher value in RMSSD and SDNN when all the participants are in rest condition and a significant drop in those values can be observed when the participants are exposed to stress simulation. This evidenced that the theory is true where a subject will have a lower HRV value during stress condition compared to rest condition.

All participants show high fluctuation in IBI values during rest condition and a low fluctuation in those values can be observed when the participants are exposed to stress simulation. A high fluctuation in IBI values will lead to a higher RMSSD and SDNN values and vice versa. From box-plot of IBI for all subjects, it can be clearly shown that all participants have bigger IQR values during rest condition compared to stress condition. A big IQR value indicates that variation in IBI values during rest is high and will lead to higher HRV values. Only Subject 1 exhibits a big difference in median values during rest and stress conditions. This indicates that Subject 1 has a high fitness level in handling stress compared to other subjects.

## 5 Conclusion

In this study the protocols from Montreal Imaging Stress Task (MIST) is able to create stress simulation in all participants where all of them are required to answer a series of computerized mental arithmetic challenges in 10 min. When participants are in rest condition, the HRV values in the RMSSD and the SDNN values are higher implying that the stress elements are absent. On the other hand, all participants show obvious drop in the HRV value when they are exposed to stress simulation. In this study, results also show that the RMSSD value is more stable and consistent in determining the HRV value compared to the SDNN.

Besides, all participants show changes from high fluctuation to low fluctuation in IBI values when they are exposed to stress simulation. Low fluctuation in IBI values will lead to a lower HRV value indicating the presence of stress elements. In this study, PPG technique is able to record the HR data accurately as showcased by the

ECG technique. The PPG sensor has advantages in terms of flexibility and portability and can compete with ECG technique in the future.

In conclusion, this study supports the theory where a subject will have a lower HRV value during stress condition compared to rest condition. However, this study has some limitations where the involved participants were chosen randomly without taking into considerations gender and lifestyle. The lifestyles of subject such as caffeine intake, smoking and physical activity are believed to have a considerable effect on the heart rate (HR).

**Acknowledgements** The authors would like to thank the Universiti Kebangsaan Malaysia for partly supporting this work through Research University Grant (GUP-2018-050).

## References

1. Page T (2015) A forecast of the adoption of wearable technology. Int J Technol Diffus 6(2):12–29
2. Kang SJ, Ha GC, Ko KJ (2017) Association between resting heart rate, metabolic syndrome and cardiorespiratory fitness in Korean male adults. J Exerc Sci Fitness 15(1):27–31
3. Jayadevappa BM, Holi MS (2007) Photoplethysmography: design, development, analysis and applications in clinical and physiological measurement—a review. Int J Innov Res Sci Eng Technol 3519–3531 (An ISO 3297(i))
4. Charlton PH, Celka P, Farukh B, Chowienczyk P, Alastruey J (2018) Assessing mental stress from the photoplethysmogram: a numerical study. Physiol Meas 39:054001
5. Elgendi M (2012) On the analysis of fingertip photoplethysmogram signals. Curr Cardiol Rev 8(1):14–25
6. Billman GE, Huikuri HV, Sacha J, Trimmel K (2015) An introduction to heart rate variability: methodological considerations and clinical applications
7. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH (2018) Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Invest 15(3):235–245
8. Butz AM (2017) 乳鼠心肌提取 HHS Public Access. Physiol Behav 176(12):139–148
9. Li F, Xu P, Zheng S, Chen W, Yan Y, Lu S, Liu Z (2018) Photoplethysmography based psychological stress detection with pulse rate variability feature differences and elastic net. Int J Distrib Sens Netwo 14(9). https://doi.org/10.1177/1550147718803298
10. Catai AM, Pastre CM, de Godoy MF, da Silva E, de Medeiros Takahashi AC, Vanderlei LCM (2020) Heart rate variability: are you using it properly? Standardisation checklist of procedures. Braz J Phys Ther 24(2):91–102
11. Wang C, Li Z, Wei X (2013) Monitoring heart and respiratory rates at radial artery based on PPG. Opt Int J Light Electron Opt 124(4):3954–3956
12. Dedovic K, Renwick R, Mahani NK, Engert V, Lupien SJ, Pruessner JC (2005) The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. J Psychiatry Neurosci 30(5):319–325
13. Vest AN, Da Poian G, Li Q, Liu C, Nemati S, Shah AJ, Clifford GD (2018) An open source benchmarked toolbox for cardiovascular waveform and interval analysis. Physiol Meas 39(10):105004
14. Bravi A, Longtin A, Seely AJE (2011) Review and classification of variability analysis techniques with clinical applications. Biomed Eng Online 10(1):90
15. Wang H, Huang S (2012) SDNN/RMSSD as a Surrogate for LF/HF: a revised investigation

# Rational Finite Difference Solution of First-Order Fredholm Integro-differential Equations via SOR Iteration


**Ming Ming Xu, Jumat Sulaiman, and Labiyana Hanif Ali**

**Abstract** The linear rational finite difference method (LRFD) is becoming more and more popular recently due to its excellent stability properties and convergence rate, especially when we are approximating the derivative of some points near the end of the interval. The main intention of this paper is to combine the 3-point linear rational finite difference (3LRFD) method with the composite trapezoidal (CT) quadrature formula to discretize the first-order linear integro-differential equation and produce dense linear systems. Furthermore, the numerical solution of the integro-differential equation is obtained by implementing the Successive Over-Relaxation (SOR) method. At the same time, the classical Gauss–Seidel (GS) method is also introduced as the control condition. In the end, through several numerical examples, the number of iterations, the execution time and the maximum absolute error are compared, which fully illustrated the superiority of SOR method over GS method in solving large dense linear system generated by the CT-3LRFD formula.

**Keywords** Integro-differential equations · First-order linear Fredholm equations · Successive over-relaxation method · Linear rational finite difference · Composite trapezoidal quadrature formula

## 1 Introduction

As mathematical models, integro-differential equations (IDEs) are widely applied in various subjects, such as mathematics, physics, chemistry, biology and geography etc.

M. M. Xu (✉)
School of Mathematics and Information Technology, Xingtai University, 88 Quanbei East Street, Xiangdu District, Xingtai, Hebei, China
e-mail: xmmzg@sina.com

M. M. Xu · J. Sulaiman · L. Hanif Ali
Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, 88400 Sabah, Malaysia
e-mail: jumat@ums.edu.my

L. Hanif Ali
e-mail: labiyana15@gmail.com



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_38

[1]. Fredholm and Volterra integro-differential equations are two classical types of IDE. However, it is difficult to find their analytic solution in many practical problems. Therefore, in the recent years of research progress, the numerical solution of this problem has attracted a great deal of attention and has been covered by a large number of papers and books, for instance, spline collocation method [2], Chebyshev and Legendre polynomials [3], the reproducing kernel method [4], operational matrix [5], CAS wavelets [6] and Taylor expansion method [7]. The research in this paper focuses on Fredholm integro-differential equation (FIDE), which is defined as follows

$$\begin{aligned} y'(t) &= p(t)y(t) + f(t) + \int_a^b K(t,u)y(u)du,\ a < t \le b, \\ y(a) &= y_a, \end{aligned} \tag{1}$$

where the functions $f(t)$, $p(t)$ and the kernel $K(t,u)$ are known and $y(t)$ is the solution to be determined.

In 2012 [8], Klein and Berrut proposed the linear rational finite difference (LRFD) methods based on the linear barycentric rational interpolants (LBRI), which can be applied to approximate the derivative of the interpolated function. Compared with the classical finite difference (FD) method, the LRFD method has better stability and accuracy, especially when it is used to approximate the derivative of the point near the end of the interval. In 2014 [9], Berrut et al. applied the linear barycentric rational quadrature method for Volterra integral equations. In 2018 [10], Abdi Hosseini applied the linear barycrntric rational method for stiff Volterra integral equations, further, in 2019, they extended LRFD method to solving stiff ODEs [11] and stiff VIEs [12]. These previous studies have motivated us to apply to the LRFD method for the FIDE problem. The main content of this paper is to apply 3-point linear rational finite difference (3LRFD) method and composite trapezoidal (CT) quadrature formula respectively to discretize the differential and integral terms of the FIDE to form a large dense linear system, and then implement SOR method to accelerate the solution process and obtain the numerical solution of problem (1).

The rest of the paper is organized along the following lines. In Sect. 2, the 3LRFD, CT and SOR methods are elaborated. In Sect. 3, we investigate three numerical examples and discuss the numerical results to verify the value of the proposed method. The last section is the conclusion of this paper.

## 2 Methodology

Since problem (1) has differential and integral terms, both terms need to be approximated in discrete form. Therefore, we mainly introduce the combination between 3LRFD scheme and CT formula, in which this combination is mainly used for discretizing problem (1). We also present in detail the SOR iterative method for solving the generated dense linear system.

### 2.1 3-Point Linear Rational Finite Difference Method

LRFD [8] methods are usually based on LBRI. Let $t_0, t_1, \cdots, t_n$ be $n+1$ real abscissas and $y(t_0), y(t_1), \ldots y(t_n)$ corresponding values. A LBRI [13] to these data will here be an expression of the form

$$Y_n(t) = \sum_{j=0}^{n} \left( \frac{\left(\frac{\xi_j}{t-t_j}\right) y(t_j)}{\left(\sum_{i=0}^{n} \frac{\xi_i}{t-t_i}\right)} \right). \tag{2}$$

In 2007, Floater and Hormann [14] gave a specific expression of weights $\xi_j$, $j = 0, 1, \cdots n$. For equispaced nodes, the weights formulas are

$$\xi_j = \frac{(-1)^{j-d}}{2^d} \sum_{k \in J_j} \binom{d}{j-k}, \; J_j := (k \in \{0, 1, 2, \cdots, n-d\} : \; j-d \le k \le j). \tag{3}$$

In this section, considering the network of uniform grids on the interval $[a, b]$ with the equispaced $h = \frac{b-a}{n}$. Refer to Eq. (2), the formula of 3LRFD to approximate the first derivative of $y(t)$ on $t_0, t_1, \ldots, t_n$ is written as

$$y'(t_i) = Y_2'(t_i) + e(t_i). \tag{4}$$

in which

$$Y_2'(t_i) = \begin{cases} \frac{1}{h} \sum_{j=i-1}^{i+1} D_{i,j}\, y(t_j), & i = 1, 2, \cdots, n-1, \\ \frac{1}{h} \sum_{j=i-2}^{i} D_{i,j}\, y(t_j), & i = n, \end{cases} \tag{5}$$

where

$$D_{i,j} = \begin{cases} \frac{\xi_j}{\xi_i}\left(\frac{1}{i-j}\right), & j \ne i, \\ -(D_{i,i-2} + D_{i,i-1}), & j = i = n, \\ -(D_{i,i-1} + D_{i,i+1}), & \text{others}. \end{cases} \tag{6}$$

In our study, we apply the 3LRFD formula to discretize the differential part of the problem (1) to derive the corresponding quadrature-rational finite difference approximation equation for the problem (1). We mainly focused on the 3LRFD at $d = 1$, and then the corresponding value of $D_{i,j}$ can be obtained from Eq. (6), as shown in Table 1, also the corresponding order of error accuracy can be obtained

**Table 1** The value of $D_{i,j}$

| $d = 1$ | $D_{i,i-1}$ | $D_{i,i}$ | $D_{i,i+1}$ |
|---|---|---|---|
| $i = 1, 2, \cdots, n-1$ | $-\frac{1}{2}$ | 0 | $\frac{1}{2}$ |
| $i = n$ | $\frac{1}{2}$ | 2 | $\frac{3}{2}$ |

from Berrut et al. [15] as $|e(t_i)| = O(h)$.

## 2.2 Quadrature Method

The integral part in Eq. (1) was discretized by applying CT formula from the family of quadrature methods to construct approximation equations coincide with the differential part. Generally, the quadrature formula can be defined as follows

$$\int_a^b y(u)du = \sum_{j=0}^{n} C_j y\left(u_j\right) + \delta_n(y). \tag{7}$$

where for $j = 0, 1, \ldots, n$, $u_j$ indicates the abscissas of the partition points of the integration interval $[a, b]$, $C_j$ are independent numerical coefficients and $\delta_n(y)$ is the truncation error. To construct the formulation of the approximation equations for the problem (1), we consider the CT method. Thus, the $C_j$ based on CT method can be shown as follows:

$$C_j = \begin{cases} \frac{1}{2}h, & j = 0, n, \\ h, & \text{others}. \end{cases} \tag{8}$$

By substituting Eqs. (5), (6), (7) and (8) into Eq. (1), a general form of the first-order quadrature-rational finite difference approximation equation can be constructed as

$$\frac{1}{h}\sum_{j=0}^{n} D_{i,j} y_j = p_i y_i + f_i + \sum_{j=0}^{n} C_j K_{i,j} y_j, \quad i = 1, 2, \cdots, n, \tag{9}$$

where $y_i = y(t_i)$, $p_i = p(t_i)$, $f_i = f(t_i)$, $K_{i,j} = K(t_i, u_j)$.

Based on the approximation Eq. (9), the corresponding linear system can be constructed which can be easily shown as

$$\tilde{M}\ \tilde{y} = \tilde{F}, \tag{10}$$

where $\tilde{M} = M^T M$, $\tilde{F} = M^T F$,

$$\tilde{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{n-3} \\ y_{n-2} \\ y_{n-1} \\ y_n \end{bmatrix}_{(n\times 1)},$$

$$F = \begin{bmatrix} f_1 + \frac{1}{2h} y_0 + C_0 K_{1,0} y_0 \\ f_2 + C_0 K_{2,0} y_0 \\ f_3 + C_0 K_{3,0} y_0 \\ f_4 + C_0 K_{4,0} y_0 \\ \vdots \\ f_{n-3} + C_0 K_{n-3,0} y_0 \\ f_{n-2} + C_0 K_{n-2,0} y_0 \\ f_{n-1} + C_0 K_{n-1,0} y_0 \\ f_n + C_0 K_{n,0} y_0 \end{bmatrix}_{(n\times 1)},$$

and $M = (M_{i,j})_{n\times n} = O - P - Q$, in which

$$O = \begin{bmatrix} 0 & \frac{1}{2h} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -\frac{1}{2h} & 0 & \frac{1}{2h} & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2h} & 0 & \frac{1}{2h} & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2h} & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{2h} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & -\frac{1}{2h} & 0 & \frac{1}{2h} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\frac{1}{2h} & 0 & \frac{1}{2h} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{2h} & -\frac{2}{h} & \frac{3}{2h} \end{bmatrix}_{(n\times n)},$$

$P = diag(p_1, p_2, \cdots, p_n)$, and $Q = (Q_{i,j})_{n\times n} = (C_j K_{i,j})_{n\times n}$. Namely

$$M_{i,j} = \begin{cases} -\frac{1}{2h} - C_i K_{i,j}, & i = 2, 3, \cdots, n-1, \quad j = i-1, \\ -p_i - C_i K_{i,j}, & i = 1, 2, 3, \cdots, n-1, \ j = i, \\ \frac{1}{2h} - C_i K_{i,j}, & i = 1, 2, 3, \cdots, n-1, \ j = i+1, \\ \frac{1}{2h} - C_i K_{i,j}, & i = n, \quad j = n-2, \\ -\frac{2}{h} - C_i K_{i,j}, & i = n, \quad j = n-1, \\ \frac{3}{2h} - C_i K_{i,j}, & i = n, \quad j = n, \\ -C_i K_{i,j}, & \text{others.} \end{cases}$$

## 2.3 Successive Over-Relaxation Method

Due to the advantage of the iteration process, SOR iterative method is an effective iterative method for solving large and dense linear systems (10). Let the matrix $\tilde{M}$ which is decomposed into

$$\tilde{M} = D - L - U, \tag{11}$$

where $D$, $-L$ and $-U$ are diagonal, strictly lower triangular and strictly upper triangular matrices respectively. Thus, the general formula for the SOR iterative method can be written as [16–19]

$$\tilde{y}^{(k+1)} = (D - \omega L)^{-1}((1-\omega)D + \omega U)\tilde{y}^{(k)} + (D - \omega L)^{-1}\omega\tilde{F}, \tag{12}$$

where $\omega$ is a weighted parameter.

The SOR iterative method is attempted to find a solution to the system of linear equations iteratively with the approximate solution to the vector $\tilde{y}$. The iterations of the method are continued until the solution is within a predetermined acceptable bound on the error. By determined the values of matrices $D$, $-L$ and $-U$ as per stated in Eq. (11), the general algorithm for SOR methods to solve the problem (1) can be described as in Algorithm 1.

**Algorithm 1**: SOR methods

a. Initializing all the parameters. Set $k = 0$, and $y^{(0)} = 0$.
b. For $k = 1, 2, 3, \ldots$ calculate

$$\tilde{y}^{(k+1)} = (D - \omega L)^{-1}((1-\omega)D + \omega U)\tilde{y}^{(k)} + (D - \omega L)^{-1}\omega\tilde{F}.$$

.

c. Convergence test. If the error of tolerance $\left\|\tilde{y}^{(k+1)} - \tilde{y}^{(k)}\right\| \leq \sigma = 10^{-10}$ is satisfied, then the numerical solution is $\tilde{y}^{(k+1)}$ and the computations stop.

d. Else, set $k = k + 1$ and go to step (b).

## 3 Results and Discussion

In order to better illustrate the advantages of applying SOR method to solve numerical solutions based on the approximation equation obtained from CT-3LRFD formula introduced in this paper, three numerical examples have been considered in our work as follows:

**Example 1** [20] Consider the linear FIDE of first-order

$$y'(t) = 1 - \frac{1}{3}t + \int_0^1 tuy(u)du, \quad 0 < t \leq 1, \tag{13}$$

with its initial condition $y(0) = 0$, and exact solution of problem (13) is $y(t) = t$.

**Example 2** [21] Consider the linear FIDE of first-order

$$y'(t) = \frac{1}{6} - \frac{1}{18}t + \int_0^1 tuy(u)du, \quad 0 < t \leq 1, \tag{14}$$

with its initial condition $y(0) = 0$, and exact solution of problem (14) is $y(t) = \frac{1}{6}t$.

**Example 3** [21] Consider the linear FIDE of first-order

$$y'(t) = \cos(t) + \frac{1}{4}t - \frac{1}{4}\int_0^{\frac{\pi}{2}} tuy(u)du, \quad 0 < t \leq 1, \tag{15}$$

with its initial condition $y(0) = 0$, and exact solution of problem (15) is $y(t) = \sin(t)$.

Here the three parameters of the number of iterations, the execution time and the maximum values of absolute errors obtained from the implementation of SOR and GS methods are taken into account. As for comparisons, the classical GS iterative method acts as the control of the comparison of numerical experiments. For the Examples 1 to 3, we have carried out a large number of numerical experiments with MATLAB software, the three parameters are compared respectively, and the corresponding results are recorded in Tables 2, 3 and 4. In particular, For the first two parameters, Figs. 1, 2 and 3 provide a more intuitive comparison of using two different iteration methods.

Refer to Tables 2, 3 and 4 for n = 1024, it can be observed that the accuracy of numerical solution of GS iteration is not accurate as the SOR iteration, this is because

**Table 2** Comparison of three parameters for two different iterative methods in Example 1

| Methods | Number of iterations | | | | |
|---|---|---|---|---|---|
| | **Mesh sizes** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 41,584 | 140,462 | 495,795 | 1,789,532 | 6,515,849 |
| SOR-3LRFD ($\omega$) | 547 (1.961007) | 1044 (1.979574) | 2001 (1.989564) | 3874 (1.994721) | 8041 (1.997494) |
| | **Execution time (s)** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 0.2088 | 0.9906 | 6.9489 | 60.8101 | 1895.7701 |
| SOR-3LRFD | 0.0028 | 0.0068 | 0.0291 | 0.1743 | 2.2537 |
| | **Maximum absolution error** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 2.3215E−05 | 5.7183E−06 | 1.2026E−06 | 7.3113E−07 | 2.0422E−06 |
| SOR-3LRFD | 2.3249E−05 | 5.8112E−06 | 1.4513E−06 | 3.6129E−07 | 9.3526E−08 |

**Table 3** Comparison of three parameters for two different iterative methods at Example 2

| Methods | Number of iterations | | | | |
|---|---|---|---|---|---|
| | **Mesh Sizes** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 37,141 | 124,500 | 435,527 | 1,555,552 | 5,594,038 |
| SOR-3LRFD ($\omega$) | 526 (1.961206) | 1013 (1.980989) | 1752 (1.989222) | 3760 (1.995121) | 7414 (1.997552) |
| | **Execution time (s)** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 0.1964 | 1.0987 | 7.1231 | 59.4149 | 1430.5265 |
| SOR-3LRFD | 0.0029 | 0.0101 | 0.0315 | 0.1626 | 2.0271 |
| | **Maximum absolution error** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 3.8377E−06 | 8.7422E−07 | 2.6501E−07 | 7.3113E−07 | 2.0421E−06 |
| SOR-3LRFD | 3.8742E−06 | 9.7089E−07 | 2.4471E−07 | 6.3816E−08 | 1.8814E−08 |

the GS iteration needs to more iteration, due to large iteration, the round up error occurs during calculation.

Based on the results obtained from Tables 2, 3 and 4, they are mainly the difference between the number of iterations and the execution time, which clearly show that the number of iterations of the SOR method has been drastically reducing compared to GS method, see Figs. 1, 2 and 3, and Table 5. At the same time, the execution time of the SOR methods is much faster compared to the GS method, see Figs. 1, 2 and 3 and Table 5. The accuracy of the SOR method is in good agreement compared with

**Table 4** Comparison of three parameters for two different iterative methods at Example 3

| Methods | Number of iterations | | | | |
|---|---|---|---|---|---|
| | **Mesh Sizes** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 25,786 | 92,868 | 340,451 | 1,255,513 | 4,629,343 |
| SOR-3LRFD ($\omega$) | 463 (1.949426) | 845 (1.973541) | 1710 (1.987108) | 3205 (1.993322) | 6317 (1.996656) |
| | **Execution time (s)** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 0.1331 | 0.8219 | 5.8326 | 47.7569 | 2513.6701 |
| SOR-3LRFD | 0.0025 | 0.0086 | 0.0291 | 0.1359 | 1.7727 |
| | **Maximum absolution error** | | | | |
| | **64** | **128** | **256** | **512** | **1024** |
| GS-3LRFD | 7.0823E−05 | 1.7668E−05 | 4.3335E−06 | 8.5428E−07 | 1.7465E−06 |
| SOR-3LRFD | 7.0842E−05 | 1.7706E−05 | 4.4253E−06 | 1.1081E−06 | 2.7851E−07 |

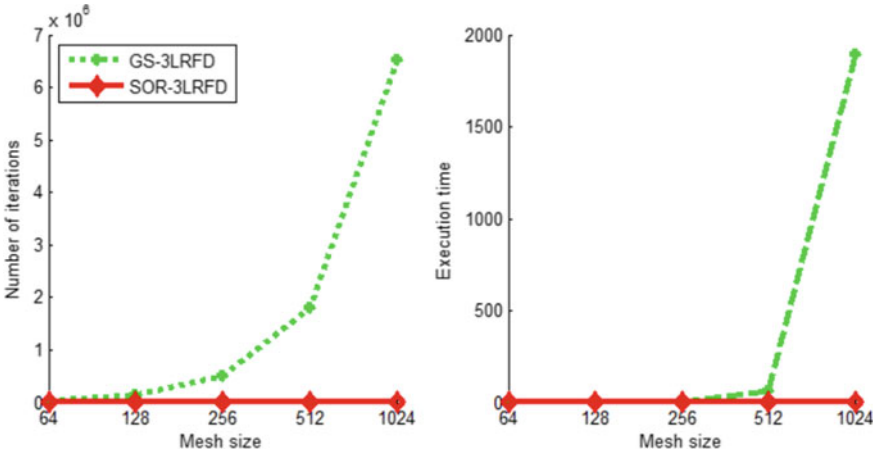



**Fig. 1** Number of iterations and execution time (s) versus mesh size of two different iteration methods in Example 1

the GS method. Overall, the advantages of the SOR method is more obvious when considering the number of iterations and the execution time.

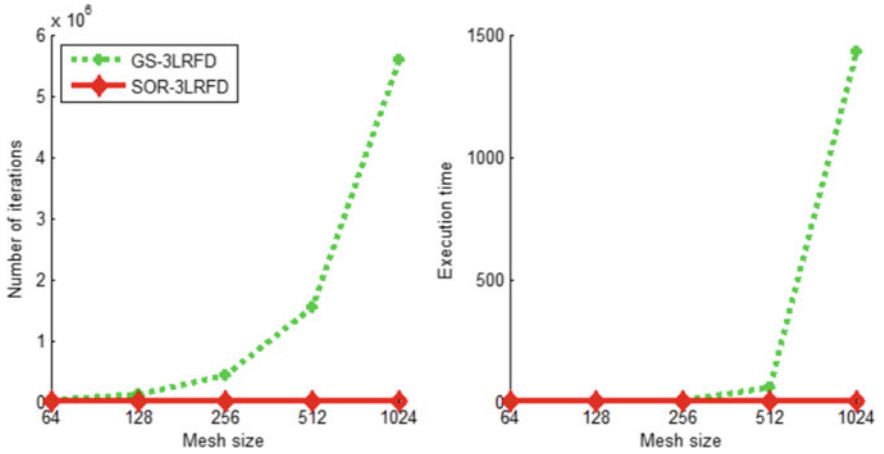



**Fig. 2** Number of iterations and execution time (s) versus mesh size of two different iteration methods in Example 2

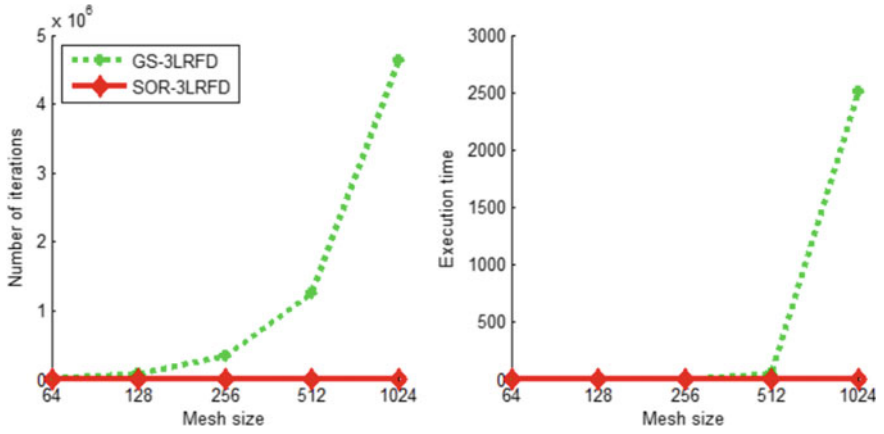



**Fig. 3** Number of iterations and execution time (s) versus mesh size of two different iteration methods in Example 3

**Table 5** Percentage reductions of number of iterations and execution time of SOR method relative to GS method in solving Examples 1 to 3 by implementing CT-3LRFD formulas

| Methods | Example | Number of iterations (%) | Execution time (%) |
|---|---|---|---|
| SOR-3LRFD | 1 | 98.68–99.88 | 98.66–9.88 |
| | 2 | 98.58–99.87 | 98.52–99.86 |
| | 3 | 98.20–99.86 | 98.12–99.93 |

## 4 Conclusion

In this paper, the 3LRFD method combined with the CT quadrature formula has been employed to discrete the differential part and integral part of the problem (1) respectively. Then SOR method is applied to accelerate the solution process so as to quickly and accurately obtain the numerical solution of the Eq. (11). Finally, in the third section, numerical experiments are used to fully demonstrate the advantages of the SOR method based on CT-3LRFD formulas used in this paper. The reason is not only that 3LRFD has good approximation property, but also that SOR iterative method has fast convergence property. In the future, we will extend the method in this paper to higher-order FIDEs problems.

## References

1. Lakshmikantham V, Rao MRM (1995) Theory of integro-differential equations, 1st edn. Gordon and Breach Science, USA
2. Dadkhah E, Shiri B, Ghaffarzadeh H, Baleanu D (2020) Visco-elastic dampers in structural buildings and numerical solution with spline collocation methods. J Appl Math Comput 63:29–57
3. Kojabad EA, Rezapour S (2017) Approximate solutions of a sum-type fractional integro-differential equation by using chebyshev and legendre polynomials. Adv Differ Equ 351:2–18
4. Jiang W, Tian T (2015) Numerical solution of nonlinear volterra integro-differential equations of fractional order by the reproducing kernel method. Appl Math Model 39:4871–4876
5. Singh VK, Postnikov EB (2013) Operational matrix approach for solution of integro-differential equations arising in theory of anomalous relaxation processes in vicinity of singular point. Appl Math Model 37:6609–6616
6. Saeedi H, Moghadam MM (2011) Numerical solution of nonlinear volterra integro-differential equations of arbitrary order by CAS wavelets. Commun Nonlinear Sci Numer Simulat 16:1216–1226
7. Huang L, Li XF, Zhao YL, Duan XY (2011) Approximate solution of fractional integro-differential equations by taylor expansion method. Comput Math Appl 62:1127–1134
8. Klein G, Berrut JP (2012) Linear rational finite differences from derivatives of barycentric rational interpolants. SIAM J Numer Anal 50(2):643–656
9. Berrut JP, Hosseini SA, Klein G (2014) The linear barycentric rational quadrature method for volterra integral equations. SIAM J Sci Comput 36:A105–A123
10. Abdi A, Hosseini SA (2018) The barycentric rational difference-quadrature scheme for systems of volterra integro-differential equations. SIAM J Sci Comput 40:A1936–A1960
11. Abdi A, Hosseini SA, Podhaisky H (2019) Adaptive linear barycentric rational finite differences method for stiff ODEs. J Comput Appl Math 357:204–214
12. Abdi A, Hosseini SA, Podhaisky H (2019) Numerical methods based on the Floater–Hormann interpolants for stiff VIEs. Numerical Algorithms, 1–20
13. Berrut JP (1988) Rational functions for guaranteed and experimentally well-conditioned global interpolation. Comput Math Appl 15(1):1–16
14. Floater MS, Hormann K (2007) Barycentric rational interpolation with no poles and high rates of approximation. Numer Math 107:315–331
15. Berrut JP, Floater MS, Klein G (2011) Convergence rates of derivatives of a family of barycentric rational interpolants. Appl Numer Math 61:989–1000

16. Akhir MKM, Othman M, Sulaiman J, Majid ZA, Suleiman M (2011) The four point-EDGMSOR iterative method for solution of 2D helmholtz equations. In: Abd Manaf A et al. (eds) ICIEIS 2011, Part III, CCIS 253. Springer, Berlin Heidelberg, pp 218–227
17. Sulaiman J, Hasan, MK, Othman M, Karim, SAA (2012) MEGSOR iterative method for the triangle element solution of 2D Poisson equations. In: International conference on computational science, ICCS 2010. Elsevier pp 377–385
18. Saudi A, Sulaiman J (2016) Path planning simulation using harmonic potential fields through four point-EDGSOR method via 9-point Laplacian. Jurnal Teknologi Sci Eng 78:12–24
19. Hong EJ, Saudi A, Sulaiman J (2017) Application of SOR iteration for poisson image blending. In: HP3C. ACM, pp 22–24
20. Darania P, Ebadian A (2007) A method for numerical Solution of integro-differential equations. Appl Math Comput 188:657–668
21. Wazwaz AM (2015) A first course in integral equations, 2nd edn. World Scientific, USA

# Semi-approximate Solution for Burgers' Equation Using SOR Iteration


**N. F. A. Zainal, J. Sulaiman, A. Saudi, and N. A. M. Ali**

**Abstract** In this article, we propose semi-approximate approach in finding a solution of Burgers' equation which is one of the partial differential equations (PDEs). Without using the Newton method for linearization, we derive the approximation equation of the proposed problem by using second-order implicit scheme together with the semi-approximate approach. Then this approximation equation leads a huge scale and sparse linear system. Having this linear system, the Successive Overrelaxation (SOR) iteration will be performed as a linear solver. The formulation and execution of SOR iteration are included in this paper. This paper proposed four examples of Burgers' equations to determine the performance of the suggested method. The test results discovered that the SOR iteration is more effective than GS iteration with less time of execution and minimum iteration numbers.

**Keywords** Burgers' equation · Semi-approximate approach · GS iteration · SOR iteration

N. F. A. Zainal (✉) · J. Sulaiman · N. A. M. Ali
Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, 88400 Sabah, Malaysia
e-mail: farah.zainal19@gmail.com

J. Sulaiman
e-mail: jumat@ums.edu.my

N. A. M. Ali
e-mail: afzamatali@yahoo.com

A. Saudi
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, 88400 Sabah, Malaysia
e-mail: azali@ums.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_39

## 1 Introduction

The nonlinear PDEs can be found in numerous areas of engineering and science [1]. Burgers' equation is one of the renown nonlinear PDEs which have been solved analytically and numerically. This equation appeared in numerous problems such as sound and shock waves in a viscous fluid, turbulence, traffic flows and nonlinear wave propagation [2]. Due mainly to the significant of this equation, various computational approaches have been developed during the last decades. These approaches generally involve finite volume, finite difference, finite element and spectral methods [3–6].

Guo et al. [3] have solved Burgers' equation by using a high-order finite volume compact scheme with a third-order strong stability preserving (SSP) Runge–Kutta scheme. Mohamed [4] presented a numerical solution for solving nonlinear Burgers' equation by using finite difference method. The numerical results obtained are compared with exact solution of Euler forward discretization (EF) and Mac Cormack discretization (MCOR). Meanwhile, Chen and Zhang [5] proposed Galerkin finite element method to approximate the solution of one-dimensional Burgers' equation.

Commonly, by imposing the finite difference discretization scheme over the nonlinear PDEs, the nonlinear approximation equation can be derived to form the corresponding nonlinear system. Then, the Newton scheme applied to the nonlinear system can develop a sequence of the corresponding linear system. To avoid of having highly computational complexity from these linear systems, we consider the formulation of the semi-approximate approach [7–9] based on the unsteady state problem being applied to convert any nonlinear system into a single linear system. Typically, the linear system can be numerically solved by using either direct methods or iterative methods in which these methods act as linear solvers. However, since the coefficient matrix of the generated linear system has sparse and huge scale, in this paper, several iterations are used to achieve the desired approximate solution of the linear system via GS and SOR iterations [10–13]. Due to the advantage of SOR iteration which is one of the effective iterations, therefore, in this paper, we mainly focus on evaluating the performance of the combination between SOR iteration and semi-approximate approach for solving the proposed problems. In addition to that, the numerical results of three examples of proposed problems have been solved numerically via the implementation of SOR iteration.

This paper includes several sections. For Sect. 2, we present the formulation of semi-approximate approach in getting the linear system. In Sect. 3 shows the formulation of the proposed iterations. Then, we tested four examples in Sect. 4 for the numerical comparison and we provide the numerical results for each example in Sect. 5. Finally, conclusion is in in the last section of this paper.

Before we go through the discretization process, let us consider the one-dimensional Burgers' equation as given in the following form [14]:

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = \gamma \frac{\partial^2 v}{\partial x^2} \tag{1}$$

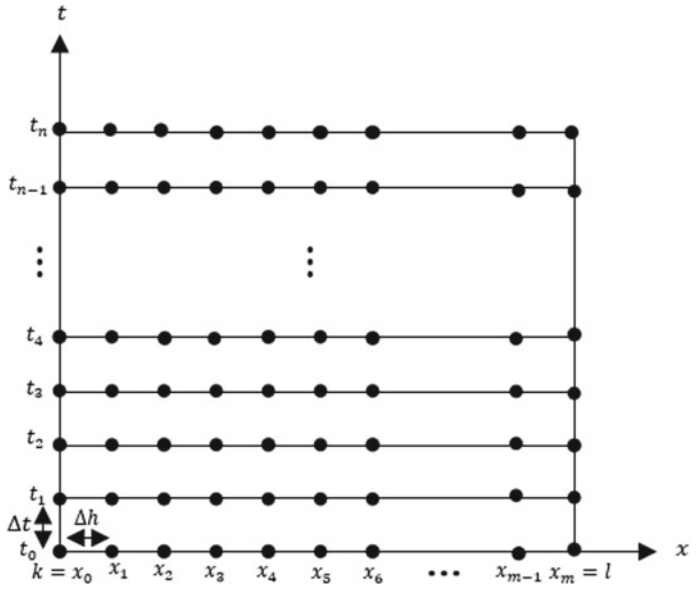



**Fig. 1** Illustration the distribution of grid points over the solution domain

with initial condition $v(x, 0) = v(x), k \leq x \leq l$ and boundary conditions $v(k, t) = f_a(t), v(l, t) = f_b(t), t > 0$. where $\gamma > 0$ is a parameter namely kinematic viscosity and $v\frac{\partial v}{\partial x}$ is the nonlinear term.

Figure 1 demonstrates the distribution of uniform grid points in which we need to construct. Then, the implementation of the GS and SOR iterative methods are implemented on interior grid points until the iterative convergence is reached. Referring to Fig. 1, let the solution domain $[k, l]$ in Eq. (1) be divided uniformly with $\Delta h$ and $\Delta t$ in directions of $x$ and $t$ respectively in which each grid of $\Delta h$ and $\Delta t$ may be described as

$$\begin{aligned} \Delta h &= \frac{l-k}{m}, \\ \Delta t &= \frac{t}{n}. \end{aligned} \tag{2}$$

## 2 Semi-approximate Approximation Equation

To start this section, the discretization of problem (1) will be presented by using the semi-approximate approach. For convenience, the nonlinear problem (1) can be simplified as

$$\frac{\partial v}{\partial t} + F(x, t, v)\frac{\partial v}{\partial x} = \lambda \frac{\partial^2 v}{\partial x^2} \tag{3}$$

To get the semi-approximate approximation equation, we impose and get the second-order implicit scheme to Eq. (3) written as (Zainal et al. [15])

$$\begin{aligned} &\frac{v_{i.j+1} - v_{i.j}}{\Delta t} + f_{i.j}\left(v_{1.j+1}, v_{2.j+1}, ..., v_{m-1.j+1}\right) \\ &= \frac{\gamma}{(\Delta h)^2}\left(v_{i-1.j+1} - 2v_{i.j+1} + v_{i+1.j+1}\right) \end{aligned} \tag{4}$$

where

$$f_{i.j}\left(v_{1.j+1}, v_{2.j+1}, ..., v_{m-1.j+1}\right) = d\left(x_i, t_{j+1}, v_{i.j+1}, \frac{v_{i+1.j+1} - v_{i-1.j+1}}{2\Delta h}\right) \tag{5}$$

Since Eq. (5) shows the nonlinear term, we solve this problem by considering the semi-approximate approach that is used to eliminate the nonlinear term of Eq. (4) in a way to form a linear system. Referring to the term $v_{i.j+1}$ in Eq. (5), several researchers have suggested to impose the semi-approximate approach in which the term $v_{i.j+1}$ will be approximated by $v_{i.j}$ as the value of $\Delta t$ is relatively small value. As a result, Eq. (5) can be approximated as

$$f_{i.j}\left(v_{1.j+1}, v_{2.j+1}, ..., v_{m-1.j+1}\right) = d\left(x_i, t_{j+1}, v_{i.j}, \frac{v_{i+1.j+1} - v_{i-1.j+1}}{2\Delta h}\right) \tag{6}$$

As provided in Eq. (4), the simplification of semi-approximate approximation equation for problem (1) can be reformulated at time level, $j + 1$ in the form of

$$-q_i v_{i-1.j+1} + r_i v_{i.j+1} - v_{i+1.j+1} = F_{i.j},\ i = 1,\ 2,\ 3,\ \ldots,\ m-1, \tag{7}$$

where

$$\begin{aligned} \beta_i &= \frac{\gamma}{(\Delta h)^2} - \left(\frac{1}{2(\Delta h)}\right)v_{i.j}, \\ q_i &= \frac{\left(\frac{1}{2(\Delta h)}\right)v_{i.j} + \frac{\gamma}{(\Delta h)^2}}{\beta_i}, \\ r_i &= \frac{1 + \frac{2\gamma\Delta t}{(\Delta h)^2}}{\beta_i}, \\ F_{i.j} &= \frac{v_{i.j}}{\Delta t.\beta_i}. \end{aligned}$$

Thus, Eq. (7) can be represented in a matrix form as

$$Bv_{j+1} = F_j \tag{8}$$

where,

$$B = \begin{bmatrix} r_1 & -1 & & & \\ -q_2 & r_2 & -1 & & \\ & -q_3 & r_3 & -1 & \\ & \ddots & \ddots & \ddots & \\ & & -q_{m-2} & r_{m-2} & -1 \\ & & & -q_{m-1} & r_{m-1} \end{bmatrix}_{(m-1)\times(m-1)},$$

$$\underline{v}_{j+1} = \left[v_{1.j+1},\, v_{2.j+1},\, v_{3.j+1},\, \ldots,\, v_{m-1.j+1}\right]^T,$$

$$\underline{F}_j = \left[F_{1.j} + q_i v_{0.j+1},\, F_{2.j},\, F_{3.j},\, \ldots,\, F_{m-2.j},\, F_{m-1.j+1} + v_{m.j+1}\right]^T$$

Clearly, it can be observed that the coefficient matrix, $B$ is sparse and huge scale.

## 3 Derivation of Successive Overrelaxation Iteration

As stated in Sect. 1, this section discusses about the iterative methods that acts as a linear solver. Since, the coefficient matrix of linear system (8) is sparse and huge scale, we consider SOR iteration to get the approximate solution of the proposed problems. In this paper, GS iteration is assigned as control method to analyze the efficiency of the proposed iteration. The improvement of the GS iteration with the use of weighted parameters, $\omega$ between the range $1 \leq \omega < 2$ can accelerate the convergence rate [16–18]. Therefore, this section attempts to discuss the formulation and execution of the modification of GS iteration known as SOR iteration. The general scheme of SOR iteration is given as

$$\underline{v}_{j+1}^{(k+1)} = (C - \omega W)^{-1} + \left[(\omega\gamma - (1 - \omega W)C)v_{i.j+1}^{(k)}\right] + (C + \omega W)^{-1}\underline{F}_j \tag{9}$$

Hence, to solve the linear system (8), the algorithm of SOR iteration may also being described in Algorithm 1.

**Algorithm 1:** SOR iteration

i. Initialize $v_{j+1}^{(0)} \leftarrow 0, \quad \varepsilon \leftarrow 10^{-10}$

ii. For $j = 1, 2, \ldots, n$, perform

a) Initialize $v_{j+1}^{(0)} = 0$.

b) Calculate vector $\underline{F_j}$.

c) For $i = 1, 2, \ldots, m-1$, compute $v_{i.j+1}^{(k+1)}$ using

$$v_{i.j+1}^{(k+1)} = (1-\omega)v_{i.j+1}^{(k)} + \frac{\omega}{r_i}\left(F_{i.j} + q v_{i-1.j+1}^{(k+1)} + v_{i+1.j+1}^{(k)}\right)$$

d) Perform the convergence test, $\left|v_{i.j+1}^{(k+1)} - v_{i.j+1}^{(k)}\right| \leq \varepsilon = 10^{-10}$. If yes, go to step e. Or else go back to step c.

e) Get the current value, $\underline{v_j}$.

iii. Display the outcome.

## 4 Numerical Examples

Considering that the primary goal for this paper is to verify the effectiveness of SOR iteration, hence, four examples of Burgers' problems are considered. This paper evaluated three aspects for comparison purpose which is iteration numbers, time of execution and maximum absolute error. We test the execution of GS and SOR iterations with different grid sizes, $m = 256, 512, 1024, 2048$ and 4096.

**Example 1** [19] Consider the following initial value equation:

$$v(x, 0) = 2x, \text{ for } t > 0. \tag{10}$$

Its exact solution is given by

$$v(x, t) = \frac{2x}{1 + 2t}. \tag{11}$$

**Example 2** [3] Consider the following initial value equation:

$$v(x, 0) = 2\gamma \frac{\pi \sin(\pi x)}{\sigma + \cos(\pi x)}, \text{ for } t > 0. \tag{12}$$

The exact solution of problem (12) is given by

$$v(x, t) = \frac{2\gamma\pi e^{-\pi^2\gamma t} \sin(\pi x)}{\sigma + e^{-\pi^2\gamma t} \cos(\pi x)}. \tag{13}$$

**Example 3** [20] Consider the following initial value equation:

$$v(x,1) = \frac{x}{1+\exp\left(\frac{1}{4\gamma}\left(x^2-\frac{1}{4}\right)\right)}, \text{ for } t > 0. \tag{14}$$

Its exact solution is given by

$$v(x,t) = \frac{\frac{x}{t}}{1+\left(\frac{t}{t_0}\right)^{\frac{1}{2}}\exp\left(\frac{x^2}{4\gamma t}\right)}, \text{ where } t_o = \exp\left(\frac{1}{8\gamma}\right) \tag{15}$$

**Example 4** [21] Consider problem (1) with initial value equation are taken from the exact solution [22].

$$v(x,t) = \frac{\lambda}{1+\lambda t}\left(x+\tan\left(\frac{x}{2+2\lambda t}\right)\right), \ t \geq 0. \tag{16}$$

## 5 Discussion

As discussed in the previous section, there are four numerical examples which are being evaluated. Tables 1, 2 and 3 shows the numerical results for iteration numbers, time of execution and maximum absolute error for examples 1, 2, 3 and 4 which can be illustrated by Figs. 2, 3, 4, 5, 6, 7, 8 and 9. Meanwhile, Table 4 indicates the reduction percentage for SOR iteration compared to GS iteration in aspects of iteration numbers and time of execution.

Based on the numerical results obtained in Tables 1, 2 and 3 which can be illustrated as shown in Figs. 2, 3, 4, 5, 6, 7, 8 and 9, it clearly shows a significant difference between GS and SOR iterations. Meanwhile, in Table 4 show the reduction percentage of iteration number and time of execution. It can be described that

**Table 1** Iteration numbers for Examples 1, 2, 3 and 4

| Example | m | 256 | 512 | 1024 | 2048 | 4096 |
|---|---|---|---|---|---|---|
| | Methods | Iteration numbers | | | | |
| 1 | GS | 9442 | 34,408 | 124,303 | 444,081 | 1,564,185 |
| | SOR | 324 | 627 | 1275 | 2442 | 4906 |
| 2 | GS | 1092 | 3986 | 14,490 | 52,196 | 185,757 |
| | SOR | 143 | 275 | 530 | 1021 | 1987 |
| 3 | GS | 113 | 390 | 1395 | 4988 | 17,651 |
| | SOR | 26 | 48 | 88 | 163 | 314 |
| 4 | GS | 30 | 89 | 304 | 1076 | 3818 |
| | SOR | 22 | 41 | 79 | 154 | 300 |

**Table 2** Time of execution for Examples 1, 2, 3 and 4

| Example | m | 256 | 512 | 1024 | 2048 | 4096 |
|---|---|---|---|---|---|---|
| | Methods | Time of execution (s) | | | | |
| 1 | GS | 9.32 | 68.08 | 494.03 | 3612.07 | 25,623.28 |
| | SOR | 0.45 | 1.41 | 5.56 | 21.40 | 90.97 |
| 2 | GS | 1.14 | 8.11 | 58.66 | 424.40 | 3115.74 |
| | SOR | 0.21 | 0.66 | 2.34 | 8.90 | 36.05 |
| 3 | GS | 0.14 | 0.83 | 5.81 | 41.50 | 298.51 |
| | SOR | 0.09 | 0.15 | 0.49 | 1.82 | 7.30 |
| 4 | GS | 0.11 | 0.22 | 1.29 | 8.51 | 61.32 |
| | SOR | 0.10 | 0.14 | 0.40 | 1.38 | 5.47 |

**Table 3** Maximum absolute error for Examples 1, 2, 3 and 4

| Example | m | 256 | 512 | 1024 | 2048 | 4096 |
|---|---|---|---|---|---|---|
| | Methods | Maximum absolute error | | | | |
| 1 | GS | 6.386E−07 | 2.554E−06 | 1.021E−05 | 4.077E−05 | 1.627E−04 |
| | SOR | 9.206E−10 | 1.454E−08 | 1.167E−08 | 8.950E−09 | 8.307E−09 |
| 2 | GS | 7.900E−04 | 7.866E−04 | 7.815E−04 | 7.634E−04 | 6.920E−04 |
| | SOR | 7.904E−04 | 7.881E−04 | 7.875E−04 | 7.873E−04 | 7.873E−04 |
| 3 | GS | 6.399E−04 | 6.299E−04 | 6.279E−04 | 6.290E−04 | 6.357E−04 |
| | SOR | 6.398E−04 | 6.298E−04 | 6.273E−04 | 6.267E−04 | 6.265E−04 |
| 4 | GS | 1.064E−08 | 4.667E−08 | 1.854E−07 | 7.604E−07 | 3.048E−06 |
| | SOR | 7.908E−08 | 7.810E−08 | 6.347E−08 | 4.534E−08 | 1.964E−08 |

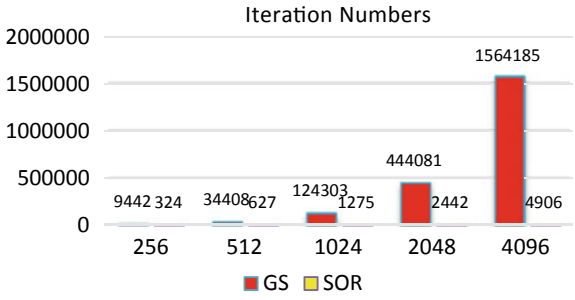



**Fig. 2** Numerical solution for iteration numbers of Example 1

the iteration numbers for SOR iteration have declined tremendously approximately by 96.57–99.67%, 86.90–98.93%, 76.99–98.22% and 26.67–92.14% respectively as compared with GS iteration. Meanwhile, for the execution time of SOR as compared with GS are more faster by 95.17–99.64%, 81.58–98.84%, 35.71–97.55% and 9.09–91.08%. Clearly, we can conclude that for all chosen grid sizes, the SOR iteration

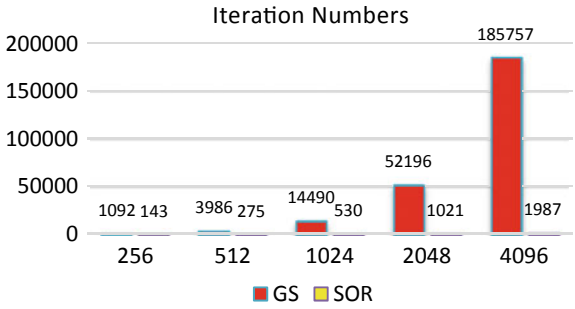



**Fig. 3** Numerical solution for iteration numbers of Example 2

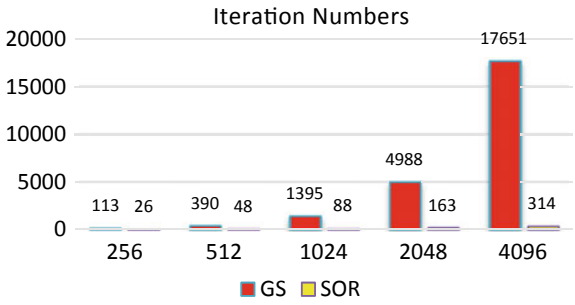



**Fig. 4** Numerical solution for iteration numbers of Example 3

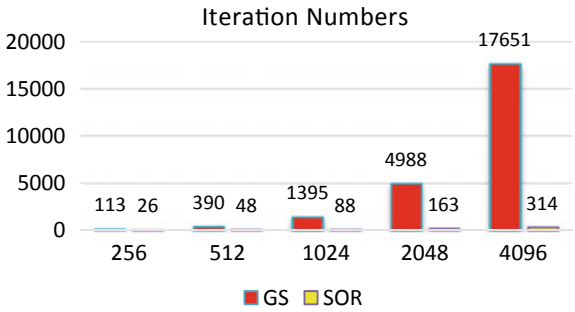



**Fig. 5** Numerical solution for iteration numbers of Example 4

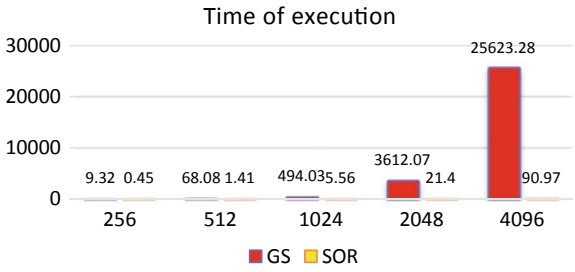



**Fig. 6** Numerical solution for time of execution of Example 1

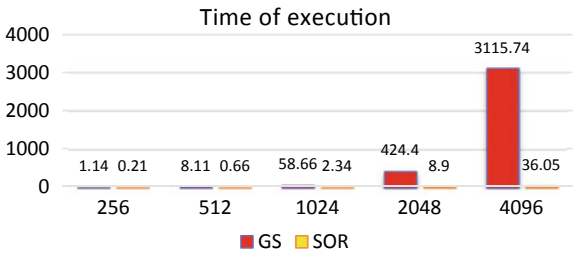



**Fig. 7** Numerical solution for time of execution of Example 2

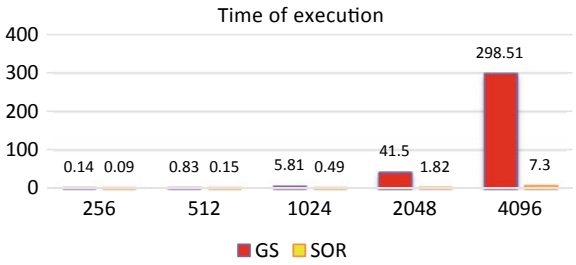



**Fig. 8** Numerical solution for time of execution of Example 3

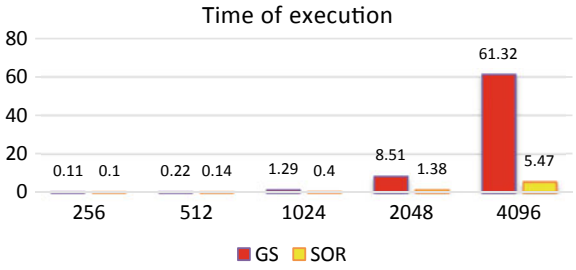



**Fig. 9** Numerical solution for time of execution of Example 4

**Table 4** Reduction percentage for SOR iteration compared to GS iteration for Examples 1, 2, 3 and 4

| Example | Iteration numbers (%) | Time of execution (%) |
|---|---|---|
| 1 | 96.57–99.67 | 95.17–99.64 |
| 2 | 86.90–98.93 | 81.58–98.84 |
| 3 | 76.99–98.22 | 35.71–97.55 |
| 4 | 26.67–92.14 | 9.09–91.08 |

produce the lowest iteration numbers and shortest time of execution relative to GS iteration.

## 6 Conclusion

As discussed in the previous section, to create a linear system of proposed problem, the second-order implicit scheme and semi-approximate approach have been successfully used to derive the corresponding second-order semi-approximate approximation equation. The iteration number and time of execution for SOR iteration is tremendously decreased as shown in Tables 1 and 2. Hence, in the end of this paper, we can conclude that the implementation of SOR iteration using semi-approximate approach is successfully solve the Burgers' equations. Overall, the numerical results have been obtained from this study is based on the point iteration family, for future work we should extend to deal with the application of block iteration family such as Explicit Group (EG) iteration [23, 24].

**Acknowledgements** The authors thankfully recognized that this paper was mainly financed by Postgraduate Centre from Universiti Malaysia Sabah, Kota Kinabalu, Malaysia.

## References

1. Kutluay S, Esen A, Dag I (2004) Numerical solutions of the Burgers' equation by the least-squares quadratic B-spline finite element method. J Comput Appl Math 167:21–33
2. Mohamed NA (2019) Solving one-and two-dimensional unsteady Burgers' equation using fully implicit finite difference schemes. Arab J Basic Appl Sci 26(1):254–268
3. Guo Y, Shi Y, Li Y (2016) A fifth-order finite volume weighted compact scheme for solving one-dimensional Burgers' equation. Appl Math Comput 281:172–185
4. Mohamed NA (2018) Fully implicit scheme for solving Burgers' equation based on finite difference method. Egypt Int J Eng Sci Tech 26:38–44
5. Chen Y, Zhang T (2019) A weak galerkin finite element method for Burgers' equation. J Comput Appl Math 348:103–119
6. Zhang Y, Lin J, Reutskiy S, Sun H, Feng W (2020) The improved backward substitution method for the simulation of time-dependent nonlinear coupled Burgers' equations. Results Phys 18
7. Rahman K, Helil N, Yimin R (2010) Some new semi-implicit finite difference schemes for numerical solution of Burgers equation. In: 2010 International conference on computer application and system modelling, vol 14, pp 451–455
8. Ozis T, Aslan Y (2005) The semi-approximate approach for solving Burgers' equation with high reynolds number. Appl Maths Comp 163:131–145
9. Liao W (2008) An implicit fourth-order compact finite-difference scheme for one-dimensional Burgers' equation. Appl Math Comput 206:755–764
10. Young DM (1954) Iterative methods for solving partial difference equations of elliptic type. Trans Am Math Soc 76(1):92–111
11. Young DM (1970) Convergence properties of the symmetric and unsymmetric successive overrelaxation methods and related methods. Math Comput 24(112):798–807
12. Sunarto A, Sulaiman J, Saudi A (2014) Full-sweep SOR iterative method to solve space-fractional diffusion equations. Aust J Basic Appl Sci 8(24):153–158
13. Zainal NFA, Sulaiman J, Alibubin MU (2018) Application of SOR iteration with nonlocal arithmetic discretization scheme for solving burger's equation. In: AIP Conference Proceedings, vol 2013, no 1, p 020035
14. Sari M, Tunc H, Seydaoglu M (2019) Higher order splitting approaches in analysis of the Burgers equation. Kuwait J Sci 46(1):1–14

15. Zainal NFA, Sulaiman J, Alibubin MU (2019) Application of half-sweep SOR iteration with nonlocal arithmetic discretization scheme for solving Burgers' equation. ARPN J Eng Appl Sci 14(3):616–621
16. Muhiddin FA, Sulaiman J, Sunarto A (2019) Grunwald implicit solution for solving one-dimensional time-fractional parabolic equations using SOR iteration. J Phys Conf Ser 1358
17. Ali NAM, Rahman R, Sulaiman J, Ghazali K (2019) Solutions of reaction-diffusion equations using similarity reduction and HSSOR iteration. Indonesian J Electr Eng Comput Sci 16:1430–1438
18. Mohamad NS, Sulaiman J (2019) The piecewise collocation solution of second kind Fredholm by using quarter-sweep iteration. J Phys Conf Ser 1358
19. Biazar J, Aminikhah H (2009) Exact and numerical solutions for non-linear Burger's equation by VIM. Math Comput Model 49:1394–1400
20. Tamsir M, Dhiman N, Srivastava VK (2016) Extended modified cubic B-spline algorithm for nonlinear Burger's equation. Beni-Suef Univ J Basic Appl Sci 5:244–254
21. Arora G, Joshi V (2017) A computational approach using modified trigonometric cubic B-spline for numerical solution of Burgers equation in one and two dimensions. Alex Eng J
22. Raslan KR (2003) A collocation solution for Burgers equation using quadratic B-spline finite elements. Int J Comput Math 80:931–938
23. Evans DJ (1985) Group explicit iterative methods for solving large linear systems. Int J Comput Math 17(1):81–108
24. Zainal NFA, Sulaiman J, Alibubin MU (2019b) Application of four-point EGSOR iteration with nonlocal arithmetic mean discretization scheme for solving Burger's equation. J Phys Conf Ser 1358:012051

# Solution of One-Dimensional Boundary Value Problem by Using Redlich-Kister Polynomial


**Mohd Norfadli Suardi and Jumat Sulaiman**

**Abstract** In this paper, the Redlich-Kister (RK) polynomial interpolation have been formulated and analyzed in solving two-point boundary value problems (BVPs). The Redlich-Kister polynomial interpolation is tested with certain number of different sizes and compared with Cubic Trigonometry B-Spline Interpolation Method (CTBIM) and Power Polynomial (Power). To do that, the discretization process of BVPs by imposing the generated RK dense linear system. Then this dense linear system need to be solved via direct method to determine the approximate value of unknown coefficients in which these coefficient are used to form the RK approximation function. Based on the maximum norm (MaxNorm) and $L^2$-Norm, the results showed that the solution by using the RK approximate function is the more accurate compared with CTBIM and Power methods.

**Keywords** Redlich-Kister · Two-point boundary value problem · Polynomial solution

## 1 Introduction

Nowadays, the numerical solutions are very important for solving many problems in various field including sciences, physics and engineering, however, the two-point BVPs are the most popular among researchers [1]. Commonly, the two-point BVPs appear in chemical modeling, heat transfer and absorption and optimal optimum issues [2]. Due to this attention, various methods have been used like Sinc-Galerkin method and modifications decomposition, Adomain decomposition method and hybrid Galerkin method [3–5] for solving the proposed problem. The shooting method, the family of spline and B-spline methods have been applied for solving the same BVPs [6–9]. To begin, consider the following two-point BVPs equation at

M. N. Suardi (✉) · J. Sulaiman
Mathematics with Economics Program, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
e-mail: norfadli1412@gmail.com

J. Sulaiman
e-mail: jumat@ums.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_40

interval $[a, b]$ [10]

$$\frac{\partial^2 y}{\partial x^2} + Z(x)\frac{\partial y}{\partial x} + G(x)y = r(x), x \in [a, b] \tag{1}$$

with the Dirichlet conditions

$$y(a) = D, \quad y(b) = E$$

where D and E are assumed as left and right boundaries respectively.

According to previous studies on RK polynomial interpolation, usually this method have been used for development of mathematical models in physics and chemistry area [11–13] but only few studies in numerical analysis have been carried out [14]. That means that this method still has not been explored and applied by other researchers in numerical analysis. In addition, many researchers also earn more interest to investigate the several numerical discretization scheme especially the finite difference method for solving the boundary value problem such as standard finite difference [15], Chebyshev finite difference [16] and Rational Finite Difference [17]. They was combined a standard numerical methods with the finite difference discretization scheme to construct a new finite difference method, for example the combination between an exponential approximation and finite difference discretization scheme are known as exponential finite difference for solving two-point boundary value problems [18]. Based on that combination concept, this paper was persuaded to propose a new finite difference known as Redlich-Kister Finite Difference (RKFD) discretization scheme by imposing into the proposed problem (1) in order to get its accurate approximate solution. To do this, let us define the RK approximation function as

$$U_n(x) = \sum_{k=0}^{n} a_k \cdot T_k(x) \tag{2}$$

where $a_k, k = 0, 1, 2, \ldots, n$ are unknown parameters that need to be determined.

Here the outline of this paper is follows: Sect. 2 discusses about the RK polynomial. Then, in Sect. 3, the numerical examples are considered and the results are given. The last section deals with the conclusion based on the results recorded.

## 2 Redlich-Kister Polynomial

As mentioned in previous section, we consider the RK function for solving two-point boundary value problems. In Sect. 1, by looking the general formula in Eq. (2) and taking n = 10, we have the 10th-order RK approximation function as follows

$$U(x) = A_0T_0(x) + A_1T_1(x) + A_2T_2(x) + A_3T_3(x) + A_4T_4(x) + A_5T_5(x) + A_6T_6(x) + A_7T_7(x) + A_8T_8(x) + A_9T_9(x) + A_{10}T_{10}(x) \quad (3)$$

where

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= x(1-x)(2x-1), \\ &\vdots \\ T_n(x) &= x(1-x)(2x-1)^{n-2}. \end{aligned}$$

Refer to Eq. (3), we need to find the first and second derivative of $U(x)$. Firstly, we have obtained the first derivative as follows

$$U'(x) = A_0T_0'(x) + A_1T_1'(x) + A_2T_2'(x) + A_3T_3'(x) + A_4T_4'(x) + A_5T_5'(x) + A_6T_6'(x) + A_7T_7'(x) + A_8T_8'(x) + A_9T_9'(x) + A_{10}T_{10}'(x) \quad (4)$$

where

$$T_i'(x) = \tfrac{\delta}{\delta x}(T_i(x)), \quad i = 0, 1, 2, \ldots, 10.$$

and the second derivative is shown in Eq. (5),

$$U''(x) = A_0T_0''(x) + A_1T_1''(x) + A_2T_2''(x) + A_3T_3''(x) + A_4T_4''(x) + A_5T_5''(x) + A_6T_6''(x) + A_7T_7''(x) + A_8T_8''(x) + A_9T_9''(x) + A_{10}T_{10}''(x) \quad (5)$$

where

$$T_i''(x) = \tfrac{\delta^2}{\delta x^2}(T_i(x)), \quad i = 0, 1, 2, \ldots, 10.$$

Then, by substituting Eqs. (3), (4) and (5) into (1) the Redlich-Kister approximation equation can be shown as

$$\begin{aligned} &A_0T_0''(x) + A_1T_1''(x) + A_2T_2''(x) + A_3T_3''(x) + A_4T_4''(x) + A_5T_5''(x) + A_6T_6''(x) \\ &+ A_7T_7''(x) + A_8T_8''(x) + A_9T_9''(x) + A_{10}T_{10}''(x) + Z(x)\big[A_0T_0'(x) + A_1T_1'(x) \\ &+ A_2T_2'(x)A_3T_3'(x) + A_4T_4'(x) + A_5T_5'(x) + A_6T_6'(x) + A_7T_7'(x)\big] \\ &+ G(x)[A_0T_0(x) + A_1T_1(x) + A_2T_2(x) + A_3T_3(x) + A_4T_4(x) + A_5T_5(x) \\ &+ A_6T_6(x) + A_7T_7(x) + A_8T_8(x) + A_9T_9(x) + A_{10}T_{10}(x)] = r(x) \end{aligned} \quad (6)$$

Having the complicated Eq. (6), we rearrange and get the Redlich-Kister approximation equation for problem (1) as follow

$$W_0(x)A_0 + W_1(x)A_1 + W_2(x)A_2 + W_3(x)A_3 + W_4(x)A_4 + W_5(x)A_5 + W_6(x)A_6 + W_7(x)A_7 + W_8(x)A_8 + W_9(x)A_9 + W_{10}(x)A_{10} = r(x) \quad (7)$$

where

$$W_i(x) = T_i''(x) + Z(x)T_i'(x) + G(x)T_i(x) \quad (8)$$

Before we proceed the next stages, the solution domain $[a, b]$ need to be divided in several sub-interval in which each of node points can be defined as $x_i = a_0 + ih$, where $h = \frac{b-a}{n}$.

By considering all node points $x_i$, $i = 0, 1, 2, \ldots, n$, we can rewrite Eq. (7) as

$$W_{j,i}A_0 + W_{j,i}A_1 + W_{j,i}A_2 + W_{j,i}A_3 + W_{j,i}A_4 + W_{j,i}A_5 + W_{j,i}A_6 + W_{j,i}A_7 + W_{j,i}A_8 + W_{j,i}A_9 + W_{j,i}A_{10} = r_i \quad (9)$$

where

$$W_j(x_i) = W_{j,i}, \quad i = j = 0, 1, 2, \ldots, n.$$

According to Eq. (9), substitute the value of $i$ from 0 to 10, we will have the RK linear system

$$W \cdot \underline{A} = \underline{R} \quad (10)$$

Here, $W$ is the coefficient matrix of order $(n + 1)$, $\underline{A}$ is an unknown vector and $\underline{R}$ is a known vector. Hence, the unknown vector can be solved by using this formula

$$\underline{A} = W^{-1} \cdot \underline{R} \quad (11)$$

## 3 Result and Discussion

In this section, three numerical examples were considered to compare the accuracy of Redlich-Kister polynomial. All these examples are tested and the numerical result has been calculated based on the maximum norm (MaxNorm) and $L^2$-norm in order to choose the best method between CTBIM, Power and RK polynomials for solving two-point BVPs. The following is the formula that have been used to calculate the MaxNorm and $L^2$-norm:

$$MaxNorm = \|S_T(x_l) - y(x_l)\|_\infty \quad (12)$$

and

$$L^2\text{-}Norm = \|S_T(x_l) - y(x_l)\|_2 = \sqrt{\sum_l [S_T(x_l) - y(x_l)]^2} \tag{13}$$

Also, during the implementation of these approaches, CTBIM is set up as a bench marking method. For comparison purpose, this method has been used to validate the Power and RK methods at one constant grid size, $n = 10$. Therefore, three examples have been tested as follows:

**Example 1** [19]
Given the two-point BVPs as

$$\frac{\partial^2 y}{\partial x^2} - \frac{\partial y}{\partial x} = -e^{(x-1)^{-1}}, \quad x \in [0, 1] \tag{14}$$

The exact solution of this Example 1 is $y(x) = x\left(1 - e^{(x-1)}\right)$.

**Example 2** [20]
We consider two-point BVPs as follows

$$\frac{\partial^2 y}{\partial x^2} + (x+1)\frac{\partial y}{\partial x} - 2y = \left(1 - x^2\right)e^{(-x)}, \quad x \in [0, 1] \tag{15}$$

where its exact solution is given by $y(x) = (x-1)e^{-x}$.

**Example 3** [21]
The two-point BVPs as follows

$$\frac{\partial^2 y}{\partial x^2} - \left(\pi^2\right)y = -2\pi^2 \sin(\pi x), \quad x \in [0, 1] \tag{16}$$

where the exact solution is stated as $y(x) = \sin(\pi x)$.

The results of CTBIM, Power and RK solutions were recorded in Tables 1, 2, 3, 4, 5, 6, 7, 8 and 9 and illustrated in Figs. 1, 2, 3, 4, 5, 6, 7, 8 and 9 for all numerical examples are considered.

Based on the results were recorded in Tables 1, 2, 3, 4, 5, 6, 7, 8 and 9 and illustrated in Figs. 1, 2, 3, 4, 5, 6, 7, 8 and 9, the accuracy of Redlich-Kister approximate solution is closed to the exact solution for all examples that were also solved by CTBIM and Power approximation function. This means that the accuracy of RK approximate function is more accurate than other approximation function. Then, this statement is supported with the comparison for all numerical methods in term of MaxNorm and $L^2$-Norm after all numerical examples were tested as shown in Table 10. From Table 10 shows that the numerical solution of RK method gives the highest accuracy compared to the CTBIM and Power methods.

**Table 1** The results of CTBIM solution for Example 1

| Node point | Exact solution | CTBIM approximate solution | Error |
|---|---|---|---|
| 0.0 | 0.0000000000 | 0.0000000000 | 0.0000E+00 |
| 0.1 | 0.0593430340 | 0.0595188042 | 1.7577E−04 |
| 0.2 | 0.1101342072 | 0.1104710968 | 3.3689E−04 |
| 0.3 | 0.1510244089 | 0.1515015679 | 4.7720E−04 |
| 0.4 | 0.1804753456 | 0.1810643603 | 5.8901E−04 |
| 0.5 | 0.1967346701 | 0.1973979616 | 6.6329E−04 |
| 0.6 | 0.1978079724 | 0.1984969225 | 6.8895E−04 |
| 0.7 | 0.1814272455 | 0.1820800082 | 6.5276E−04 |
| 0.8 | 0.1450153975 | 0.1455543468 | 5.3895E−04 |
| 0.9 | 0.0856463238 | 0.0859750847 | 3.2876E−04 |
| 1.0 | 0.0000000000 | 0.0000000000 | 0.0000E+00 |

**Table 2** The results of power solution for Example 1

| Node point | Exact solution | Power approximate solution | Error |
|---|---|---|---|
| 0.0 | 0.0000000000 | 0.0000000000 | 0.0000E+00 |
| 0.1 | 0.0593430340 | 0.0593430339 | −1.4748E−10 |
| 0.2 | 0.1101342072 | 0.1101342070 | −1.3674E−10 |
| 0.3 | 0.1510244089 | 0.1510244088 | −1.0416E−10 |
| 0.4 | 0.1804753456 | 0.1804753456 | 8.2804E−11 |
| 0.5 | 0.1967346701 | 0.1967346708 | 6.8971E−10 |
| 0.6 | 0.1978079724 | 0.1978079747 | 2.3023E−09 |
| 0.7 | 0.1814272455 | 0.1814272515 | 5.9747E−09 |
| 0.8 | 0.1450153975 | 0.1450154110 | 1.3511E−08 |
| 0.9 | 0.0856463238 | 0.0856463517 | 2.7925E−08 |
| 1.0 | 0.0000000000 | 0.0000000540 | 5.3962E−08 |

## 4 Conclusion

In this paper, we have formulated and applied the Redlich-Kister approximation function in solving two-point BVPs. Also, we already calculated the accuracy for all approximation function are considered compared between them. Finally, from the numerical experiments, it can be concluded the RK method is more accurate among the CTBIM and Power approximation function. For future works, this paper can be extended by increasing the number of sizes and applying iterative methods [8, 10]. Moreover, this method also can be modify and combine it with trigonometry function in order to solve one dimensional boundary value problems.

**Table 3** The results of RK solution for Example 1

| Node point | Exact solution | RK approximate solution | Error |
|---|---|---|---|
| 0.0 | 0.0000000000 | 0.0000000000 | 1.3491E−15 |
| 0.1 | 0.059343034 | 0.0593430339 | −1.4170E−10 |
| 0.2 | 0.1101342072 | 0.1101342071 | −1.2465E−10 |
| 0.3 | 0.1510244089 | 0.1510244087 | −1.1591E−10 |
| 0.4 | 0.1804753456 | 0.1804753455 | −9.9036E−11 |
| 0.5 | 0.1967346701 | 0.1967346701 | −8.9702E−11 |
| 0.6 | 0.1978079724 | 0.1978079723 | −7.4620E−11 |
| 0.7 | 0.1814272455 | 0.1814272455 | −5.3137E−11 |
| 0.8 | 0.1450153975 | 0.1450153975 | −3.8882E−11 |
| 0.9 | 0.0856463238 | 0.0856463238 | −1.5915E−11 |
| 1.0 | 0.0000000000 | −0.0000000000 | −1.3491E−15 |

**Table 4** The results of CTBIM solution for Example 2

| Node point | Exact solution | CTBIM approximate solution | Error |
|---|---|---|---|
| 0.0 | −1.0000000000 | −1.0000000000 | 0.0000E+00 |
| 0.1 | −0.8140614971 | −0.8143536762 | 2.9218E−04 |
| 0.2 | −0.6545160011 | −0.6549846025 | 4.6860E−04 |
| 0.3 | −0.5180180080 | −0.5185727545 | 5.5475E−04 |
| 0.4 | −0.4016201233 | −0.4021920276 | 5.7190E−04 |
| 0.5 | −0.3027277157 | −0.3032653299 | 5.3761E−04 |
| 0.6 | −0.2190585503 | −0.2195246544 | 4.6610E−04 |
| 0.7 | −0.1486068703 | −0.1489755911 | 3.6872E−04 |
| 0.8 | −0.0896114535 | −0.0898657928 | 2.5434E−04 |
| 0.9 | −0.0405272181 | −0.0406569660 | 1.2975E−04 |
| 1.0 | 0.0000000000 | 0.0000000000 | 0.0000E+00 |

**Table 5** The results of power solution for Example 2

| Node point | Exact solution | Power approximate solution | Error |
|---|---|---|---|
| 0.0 | −1.0000000000 | −1.0000000000 | 7.5495E−15 |
| 0.1 | −0.8143536762 | −0.8143536763 | −1.0962E−10 |
| 0.2 | −0.6549846025 | −0.6549846025 | −8.0904E−11 |
| 0.3 | −0.5185727545 | −0.5185727545 | −6.6098E−11 |
| 0.4 | −0.4021920276 | −0.4021920277 | −5.0802E−11 |
| 0.5 | −0.3032653299 | −0.3032653299 | −3.8617E−11 |
| 0.6 | −0.2195246544 | −0.2195246544 | 1.5497E−11 |
| 0.7 | −0.1489755911 | −0.1489755909 | 1.8812E−10 |
| 0.8 | −0.0898657928 | −0.0898657922 | 6.2068E−10 |
| 0.9 | −0.0406569660 | −0.0406569644 | 1.5913E−09 |
| 1.0 | 0.0000000000 | 0.0000000035 | 3.4918E−09 |

**Table 6** The results of RK solution for Example 2

| Node point | Exact solution | RK approximate solution | Error |
|---|---|---|---|
| 0.0 | −1.0000000000 | −1.0000000000 | 3.2641E−14 |
| 0.1 | −0.8143536762 | −0.8143536763 | −1.1132E−10 |
| 0.2 | −0.6549846025 | −0.6549846025 | −8.1135E−11 |
| 0.3 | −0.5185727545 | −0.5185727545 | −6.3365E−11 |
| 0.4 | −0.4021920276 | −0.4021920277 | −4.3195E−11 |
| 0.5 | −0.3032653299 | −0.3032653299 | −3.4014E−11 |
| 0.6 | −0.2195246544 | −0.2195246545 | −2.2977E−11 |
| 0.7 | −0.1489755911 | −0.1489755911 | −1.0192E−11 |
| 0.8 | −0.0898657928 | −0.0898657928 | −6.8940E−12 |
| 0.9 | −0.0406569660 | −0.0406569660 | 1.4264E−12 |
| 1.0 | 0.0000000000 | −0.0000000000 | −6.6613E−15 |

**Table 7** The results of CTBIM solution for Example 3

| Node point | Exact solution | CTBIM approximate solution | Error |
|---|---|---|---|
| 0.0 | 0.0000000000 | −0.0000000000 | −3.0670E−15 |
| 0.1 | 0.3090169944 | 0.3080602147 | −9.5678E−04 |
| 0.2 | 0.5877852523 | 0.5859653492 | −1.8199E−03 |
| 0.3 | 0.8090169944 | 0.8065121127 | −2.5049E−03 |
| 0.4 | 0.9510565163 | 0.9481118513 | −2.9447E−03 |
| 0.5 | 1.0000000000 | 0.9969037960 | −3.0962E−03 |
| 0.6 | 0.9510565163 | 0.9481118513 | −2.9447E−03 |
| 0.7 | 0.8090169944 | 0.8065121127 | −2.5049E−03 |
| 0.8 | 0.5877852523 | 0.5859653492 | −1.8199E−03 |
| 0.9 | 0.3090169944 | 0.3080602147 | −9.5678E−04 |
| 1.0 | 0.0000000000 | 0.0000000000 | 0.0000E+00 |

**Table 8** The results of power solution for Example 3

| Node point | Exact solution | Power approximate solution | Error |
|---|---|---|---|
| 0.0 | 0.0000000000 | 0.0000000002 | 2.1160E−10 |
| 0.1 | 0.3090169944 | 0.3090155121 | −1.4823E−06 |
| 0.2 | 0.5877852523 | 0.5877842444 | −1.0079E−06 |
| 0.3 | 0.8090169944 | 0.8090162370 | −7.5735E−07 |
| 0.4 | 0.9510565163 | 0.9510560274 | −4.8893E−07 |
| 0.5 | 1.0000000000 | 0.9999996113 | −3.8866E−07 |
| 0.6 | 0.9510565163 | 0.9510562521 | −2.6416E−07 |
| 0.7 | 0.8090169944 | 0.8090168901 | −1.0431E−07 |
| 0.8 | 0.5877852523 | 0.5877851793 | −7.2995E−08 |
| 0.9 | 0.3090169944 | 0.3090170380 | 4.3672E−08 |
| 1.0 | 0.0000000000 | 0.0000000462 | 4.6203E−08 |

**Table 9** The results of RK solution for Example 3

| Node point | Exact solution | RK approximate solution | Error |
|---|---|---|---|
| 0.0 | 0.0000000000 | 0.0000000002 | 2.1111E−10 |
| 0.1 | 0.3090169944 | 0.3090155121 | −1.4823E−06 |
| 0.2 | 0.5877852523 | 0.5877842444 | −1.0079E−06 |
| 0.3 | 0.8090169944 | 0.8090162367 | −7.5769E−07 |
| 0.4 | 0.9510565163 | 0.9510560262 | −4.9011E−07 |
| 0.5 | 1.0000000000 | 0.9999996083 | −3.9173E−07 |
| 0.6 | 0.9510565163 | 0.9510562456 | −2.7071E−07 |
| 0.7 | 0.8090169944 | 0.8090168780 | −1.1642E−07 |
| 0.8 | 0.5877852523 | 0.5877851592 | −9.3122E−08 |
| 0.9 | 0.3090169944 | 0.3090170070 | 1.2632E−08 |
| 1.0 | 0.0000000000 | 0.0000000003 | 2.9793E−10 |

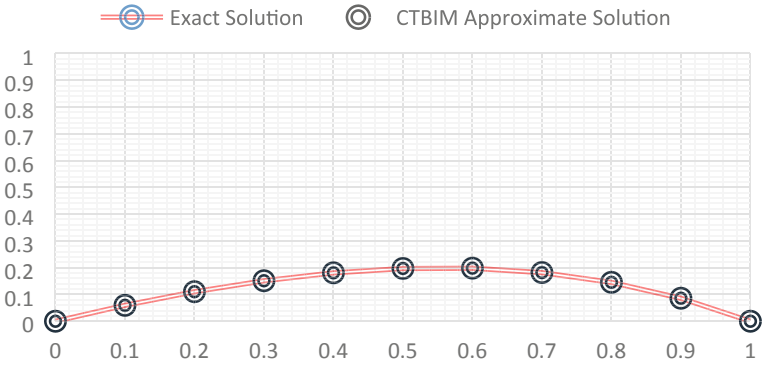



**Fig. 1** The graph of comparison CTBIM solutions for Example 1

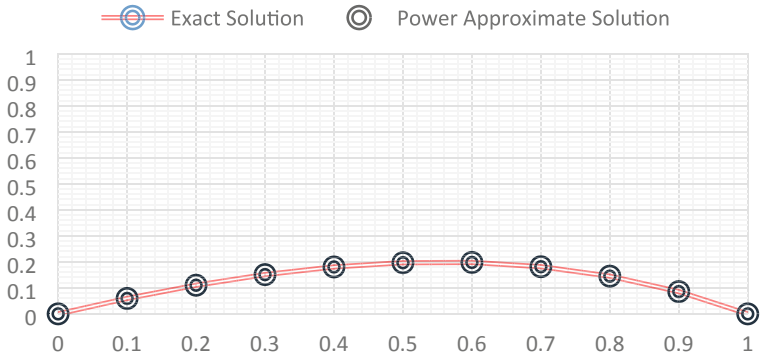



**Fig. 2** The graph of comparison Power solutions for Example 1

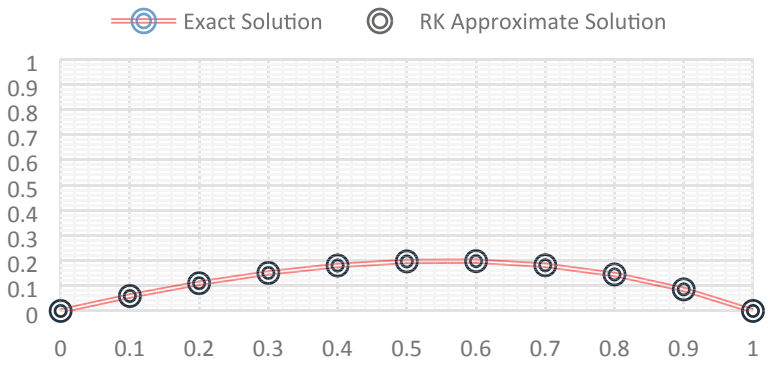



**Fig. 3** The graph of comparison RK solutions for Example 1

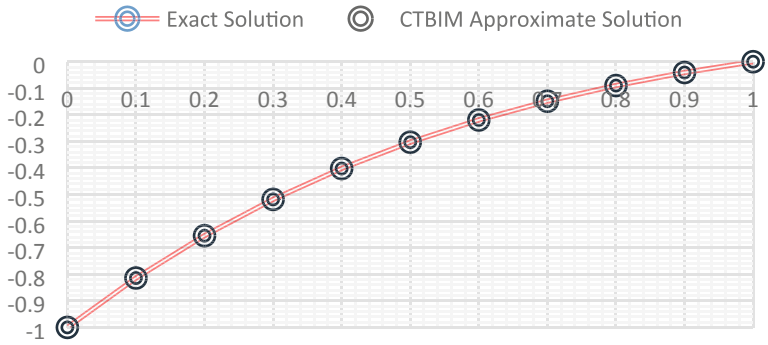



**Fig. 4** The graph of comparison CTBIM solutions for Example 2

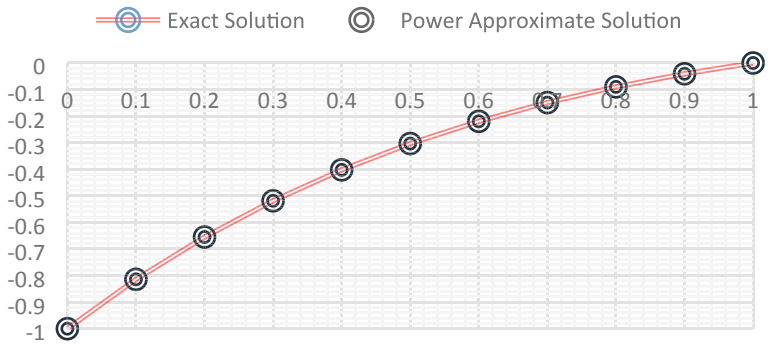



**Fig. 5** The graph of comparison power solutions for Example 2

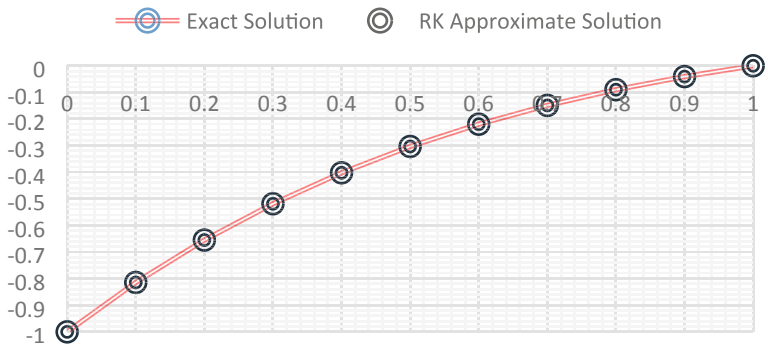



**Fig. 6** The graph of comparison RK solutions for Example 2

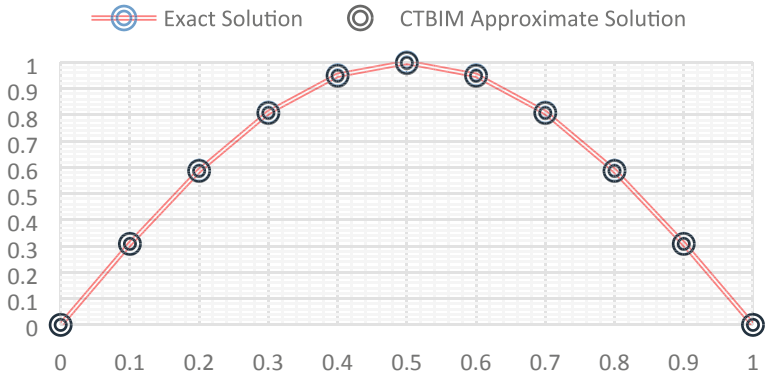



**Fig. 7** The graph of comparison CTBIM solutions for Example 3

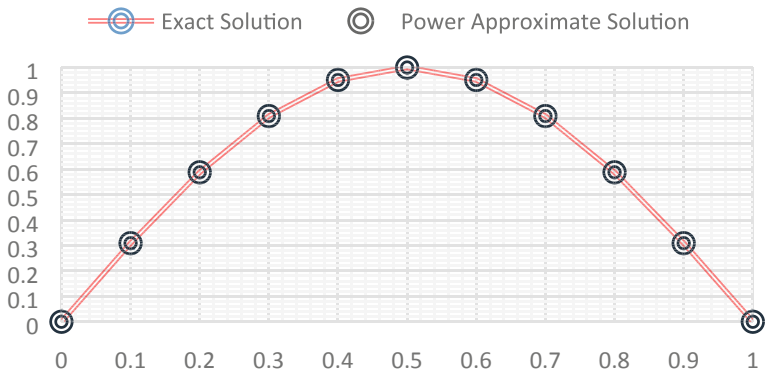



**Fig. 8** The graph of comparison power solutions for Example 3

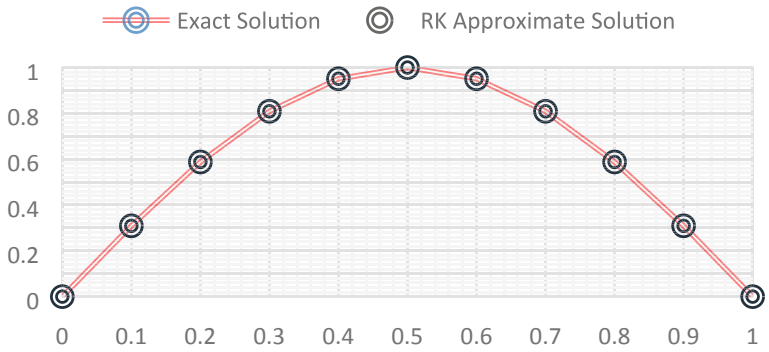



**Fig. 9** The graph of comparison RK solutions for Example 3

**Table 10** The comparison different methods for problems that are considered

| Example | Method | MaxNorm | $L^2$-Norm |
|---|---|---|---|
| 1 | CTBIM | 6.8895E−04 | 1.5679E−03 |
| | Power | 5.3962E−08 | 6.2576E−08 |
| | RK | 1.4170E−10 | 2.7760E−10 |
| 2 | CTBIM | 5.7190E−04 | 1.2898E−03 |
| | Power | 3.4918E−09 | 3.8952E−09 |
| | RK | 1.1132E−10 | 1.6339E−10 |
| 3 | CTBIM | 3.0962E−03 | 6.9233E−03 |
| | Power | 1.4823E−06 | 2.0656E−06 |
| | RK | 1.4823E−06 | 2.0680E−06 |

## References

1. Aarao J, Bradshaw-Hajek BH, Miklavcic SJ, Ward DA (2010) The extended domain eigenfunction method for solving elliptic boundary value problems with annular domains. J Phys A Math Theor 43:185–202
2. Jain PC, Kadalbajoo MK (1980) A numerical technique for solving nonlinear elliptic problems. Indian J Pure Appl Math 11(1):20–32
3. El-Gamel M (2007) Comparison of the solution obtained by Adomian decomposition and wavelet-Galerkin methods of boundary-value problems. Appl Math Comput 186(1):652–664
4. Jang B (2007) Two-point boundary value problems by extended Adomian decomposition method. Comput Appl Math 219(1):253–262
5. Mohsen A, Gamel ME (2008) On the Galerkin and collocation methods for two point boundary value problems using Sinc bases. Comput Math Appl 56:930–941
6. Lin Y, Enszer JA, Stadtherr MA (2008) Enclosing all solutions of two-point boundary value problems for ODEs. Comput Chem Eng 32(8):1714–1725
7. Nazan C, Hikmet C (2008) B-spline methods for solving linear system of second order boundary value problems. Comput Math Appl 57(5):757–762
8. Suardi MN, Radzuan NZFM, Sulaiman J (2017a) Cubic B-spline solution for two-point boundary value problem with AOR iterative method. J Phys Conf Ser 890:012015

9. Suardi MN, Radzuan NZFM, Sulaiman J (2019) Performance of quarter-sweep SOR iteration with cubic B-spline scheme for solving two-point boundary value problems. J Eng Appl Sci 14(3):693–700
10. Suardi MN, Radzuan NZFM, Sulaiman J (2017b) Cubic B-spline solution of two-point boundary value problem using HSKSOR iteration. Glob J Pure Appl Math 13(11):7921–7934
11. Babu S, Trabelsi R, Srinivasa Krishna T, Ouerfelli N, Toumi A (2019) Reduced Redlich–Kister functions and interaction studies of Dehpa + Petrofin binary mixtures at 298.15 K. Phys Chem Liq 57(4):536–546
12. Gayathri A, Venugopal T, Venkatramanan K (2019) Redlich-Kister coefficients on the analysis of physico-chemical characteristics of functional polymers. Mater Today Proc 17:2083–2087
13. Komninos NP, Rogdakis ED (2020) Geometric investigation of the three-coefficient Redlich-Kister expansion global phase diagram for binary mixtures. Fluid Phase Equilib 112728
14. Hasan MK, Sulaiman J, Ahmad S, Othman M, Abdul Karim SA (2010) Approximation of iteration number for Gauss-Seidel using Redlich-Kister polynomial. Am J Appl Sci 7:956–962
15. Chawla MM, Katti CP (1979) Finite difference methods for two-point boundary value problems involving high order differential equations. BIT Numer Math 19(1):27–33
16. Elbarbary EM, El-Kady M (2003) Chebyshev finite difference approximation for the boundary value problems. Appl Math Comput 139(2–3):513–523
17. Pandey PK (2014) Rational finite difference approximation of high order accuracy for nonlinear two point boundary value problems. Sains Malaysiana 43(7):1105–1108
18. Pandey PK (2016) Solving two point boundary value problems for ordinary differential equations using exponential finite difference method. Bol Soc Parana Mat 34(1):45–52
19. Caglar HN, Caglar SH, Elfaituri K (2006) B-spline interpolation compared with finite difference, finite element and finite volume methods which applied to two point boundary value problems. Appl Math Comput 175(1):72–79
20. Asaithambi NS (1995) Numerical analysis, theory and practice. Sauders College Publishing, Orlando, pp 361–373, 175(1), 72–79, (1995)
21. Burden RL, Faires JD (2005) Numerical analysis. Brooke/Cole, Belmont, pp 675

# Issues and Challenges for Teaching Successful Programming Courses at National Secondary Schools of Malaysia


**Faridah Hani Mohamed Salleh, Deshinta Arrova Dewi, and Nurul Azlin Liyana**

**Abstract** Undoubtedly, the initiative of the Malaysia education ministry to introduce coding in school curricula is a very good effort with lots of advantages. Able to code will be an advantage and a necessity when the students join the workforce. After not small amount of budget and time has been spent by the country for this mission, there are several issues that need to be considered and worked on to ensure coding lessons in schools achieves the target. This paper presents six issues and the recommended solutions that do not require a change to the current educations system as frequent changes in national education policy will burden teachers, parents and students. The identified issues are from the perspective of language of communication, implementation and execution, digital divide, quality tools, assessments and teaching and learning time. This works suggests to create a self-learning system built specifically for the national secondary schools' syllabus, short-term job exemption for teachers and programming skill test to replace project as part of assessments to increase the rate of teaching effectiveness of Computer Science subjects, especially programming. This study will be of great important to educational planners, school authorities, educational researchers and the governments.

**Keywords** Education · Programming · Secondary schools

F. H. M. Salleh (✉)
Department of Computing, College of Computing & Informatics, Universit Tenaga Nasional, Jalan IKRAM-UNITEN, Kajang, Selangor, Malaysia
e-mail: faridahh@uniten.edu.my

D. A. Dewi
INTI International University, Persiaran Perdana BBN, Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
e-mail: deshinta.ad@newinti.edu.my

N. A. Liyana
College of Graduate Studies, Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, Kajang, Selangor, Malaysia
e-mail: azlinliyana@yahoo.com.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_41

## 1 Introduction

The vision of our country lies in the hands of our youths. The primary role of young people is to get a good education in order to become better citizens of tomorrow. They need to learn skills to do the job that their country's economy needs. With the advent of IR 4.0, our country needs experts in Artificial Intelligence and IT-related disciplines. Programming is one of the essential skills that one need to master in if they want to be expert in IT-related disciplines. If we are able to produce high-skilled students in programming, the country will be able to produce many system makers, while reducing the dependency on outsiders. This research aims to identify the issues and challenges in teaching and learning coding among secondary school students of national schools in Malaysia. Any proposal to change the current setting of teaching programming in national schools will be avoided as frequent change of education system may not be preferred by both the government and citizens. As, such, this study will be based on the Malaysia Education Development Plan (MEDP). MEDP contains five system aspirations and six student characteristics that our country plan to achieve over the next 13 years, from 2013 to 2025. It is a comprehensive manifestation of government transformation for students from pre-school to university. Apart from educational planners, school authorities and the governments, the findings of this research will be useful to the researchers that plan to conduct research in teaching and learning coding for school students.

## 2 Coding in National Schools

Programming is an essential language to know in digital age and being able to code helps to understand so much more about all the technology. Coding fosters logical thinking and problem-solving skill. Introducing coding into school curriculum is not something new for Malaysia [1]. Programming has been introduced in Malaysia schools since 2016, starting with year six primary school. There are several studies that have been conducted previously related to programming teaching such as works by Kanemune et al. [2], Tundjungsari [3], Sklirou [4] and Jawawi et al. [5].

### *2.1 Primary School (Year Six)*

In a Malaysia public education system, year six of primary school is for the student age 12 years old. In this stage, Information and Communication Technology (ICT) is taught as a subject for preparation for high school. At this point, focus is given to mastery of knowledge and skills that fit the student's level of ability by introducing five modules; computer world, multimedia exploration, networking systems, Internet, database and programming. Scratch is being taught as a practical skill.

## 2.2 Secondary School (Upper and Lower Secondary)

In 2017, under the new Secondary School Standards-Based Curriculum (KSSM) for lower secondary students, the ministry introduced two subjects called Basic Computer Science and Design and Technology [1]. Students are given the option to learn either one of the subjects. Design and Technology subject includes topics such as product design, agriculture technology such as aquaponic, fashion and basic carpentry. Normally, only schools with computer lab facilities and IT teachers will offer Basic Science Computer subject. In Basic Computer Science, the students will be exposed to coding in different programming languages such as Scratch, HTML and Python in problem solving and projects. In the previous curriculum, when Information and Communication Technology subject was introduced, this subject emphasizes on computing only. Lower secondary is for the students age 13–15. The curriculum of the Basic Computer Science subject offered to the lower secondary students focuses on providing the students with computational thinking.

Upper secondary is for the students age 16 and 17. The fourth-form and fifth-form students can further study coding in subjects such as computer science, invention or engineering and vocational related subjects as elective subjects. Figure 1 shows level

<table>
<tr>
<td rowspan="2">Primary School: Year Six<br>Subject: Information and Communication Technology (ICT). 33% of it covers coding and programming [1].<br>Topics:<br>o Understanding programming<br>o Using algorithm through pseudo-code and flowchart<br>o Coding and debugging<br>o Programming project<br>Refer [6] and [7] for the complete syllabus.<br>Programming language/tool used: Scratch programming</td>
<td colspan="2">Secondary School<br>Student have the option to study coding in subject such as computer science, invention or engineering and vocational related subjects as relative subjects.</td>
</tr>
<tr>
<td>Lower Secondary<br>Subject: Basic Computer Science. 63% covers coding and programming [1].<br>Form One<br>Topics:<br>o Basic computational thinking<br>o Binary number system<br>o Algorithm construction<br>o Instruction codes<br>Refer [8] and [9] for the complete syllabus.<br>Programming language/ tool used: Scratch and HTML<br>Form Two<br>Topics:<br>o Data representation<br>o Algorithm<br>o Instruction codes<br>Refer [10] for the complete syllabus.<br>Programming language/ tool used: Python and Scratch<br>Form Three<br>Topics:<br>o Basic concept computational thinking<br>o Data representation<br>o Algorithm<br>o Instruction codes<br>Refer [11] for the complete syllabus.<br>Programming language/ tool used: SQL, Python</td>
<td>Upper Secondary<br>Subject: Computer Science 83% covers coding and programming [1].<br>Form Four<br>Topics:<br>o Programming<br>o Database<br>o Human interaction with computer<br>Refer [12] and [13] for the complete syllabus.<br>Programming language/technical concepts/ tool used: Systems design (ERD), SQL and database management, Microsoft Access<br>Form Five<br>o Computing<br>o Advanced database<br>o Web-based programming<br>Refer [14] for the complete syllabus<br>Programming language/technical concepts/ tool used: System design (ERD), Advanced SQL and database management, PHP myAdmin, HTML, CSS, JavaScript, data structure, PHP</td>
</tr>
</table>

**Fig. 1** Coding in national schools

and name of subject which coding is taught in Malaysia national schools. In Fig. 1 it can be seen that the second-form and third-form have continuity of learning by starting Python at the beginning and ending it with the same language before sitting for the PT3, which is one of the major examinations in Malaysia education system. Fourth-form and fifth-form emphasize the database management system. It can be seen that the emphasis on mastering programming has decreased upon the completion of third-form because the focus of the syllabus has shifted to database management.

## 3 Issues and Challenges and the Suggested Possible Solutions

There are 6 possible challenges of both teaching and learning programming in Malaysia schools have been identified. The challenges Malaysia is facing may be different from other countries in this world due to the culture differences and economic situation.

### *3.1 Language of Communication*

All the debates raging round for whether Mathematics and Science should be taught in English or Malaysia Language (Bahasa Malaysia) is applied to the teaching of Computer Science subject as well [6]. While Malaysia Language is believed to be able to increase understanding of majority of the students and preserving heritage, using English on the other hand will give more advantages such as global application and wide access to external materials.

Currently, most of the national schools use Malaysia/Malay Language in school for teaching Computer Science subject. While teaching the theories of Computer Science in the students' native language may not disclose any flaws, the students will start to show the signs of unexciting when they are required to search for additional information pertaining to programming by themselves. The difference between language used in schools and the language used later when to search for information has caused some problems, because many of the external online resources that can be used to help the students learn coding are in English. This issue becomes more prominent due to the gap in English proficiency between urban and rural students. Ibrahim et al. [7] and Maros et al. [8] reported that despite going through the same curriculum, the level of English proficiency in rural schools is much lower than the level in the urban schools. Referring to an example of a scenario related to the mother tongue, a study conducted by Ibrahim et al. [7] proves that a large number of errors identified in the English assignments used as the testing materials due to mother tongue interference. The study by Ibrahim et al. [7] was conducted among young Malay learners in Malaysian secondary schools.

Although the language has been said as one of the possible reasons of teaching Computer Science may not as efficient as what it should be, there is no published research specifically conducted has been found so far to support this fact. Thus, a survey needs to be conducted to identify whether the usage of English really cause a problem or not. And if yes, at which particular aspects really bothering the teachers and students? Does it relate to teaching resources, medium of communication, or learning resources? This is important as if the language is not the main problem, we can maintain the existing teaching methods so that efforts can be focused on improving other aspects. With the emergence of social media platforms, with many students have accessed to it, will the exposure to the "computer language" is enough for the students to at least identity the learning resources by themselves? What is the relation between ability to identify the right learning resources and English language? The answer is, the secondary school students sometimes become clueless on the choice of keywords to be used when to search for resources using search engine. They are not sure or cannot even think of any suitable keywords to be used in searching. When too many irrelevant information displayed after the arbitrary searching, they will waste too much time in filtering for the right information that suits to their needs. To conclude what has been discussed before, it is important for the country to identify whether it is correct that language is a barrier to effective programming learning, especially for those who are interested in self-study.

### *3.2 Implementation and Execution*

The detail implementation of ICT teaching and learning in for primary schools is described in [3] and in [9] for upper and secondary schools level. The desire of the Ministry of Education Malaysia to produce competitive students to face the IR 4.0 era is indeed commendable. However, there are some issues that need to be identified to ensure that the original planning of the country is in line with what has been presented in the education plan. Ideally, if the teaching of Computer Science involves a practical class, the ratio between teacher and students shall be around less than 30 students per teacher. This ratio is suitable for university undergraduate level program. However, for school level, the number of students per lab session shall be less than 20 students. This recommendation is based on the authors' experience of conducting programming lab for more than 10 years. From a study conducted by Olanrewaju and Oluyomi [10] to 150 students taking Physics in one of the secondary schools in Nigeria, it is recommended that stakeholders should put more effort into ensure that the class size is reduced to teacher-students ratio of 1:35. Any subject that requires hands-on skills does require a small class size to give teachers the opportunity to focus on students with different levels of learning mastery. Since learning to program is difficult [11] due to its nature that involves correctness of logic, syntax, and semantics; programming is a subject that is definitely included in the subject group that requires a hands-on approach.

Based on the guidelines from the Town and Country Planning Department of Selangor, the ideal class size for primary school was 30 while the ideal class size for secondary schools was 25 students per classroom [7]. This is in contrast to the current situation, where based on the findings of 3 secondary schools in Gombak district of Malaysia conducted by Ibrahim et al. [7] the class size is in the range between 29 and 38 students per class. However, this large class size may occur in schools located in densely populated areas in major Malaysian cities such as Gombak.

For programming subjects, the problem of class size seems to be easier to overcome by arranging teaching schedules. A more difficult issue here is the lack of skilled teachers in the field of programming. Class size reduction is a popular but expensive educational reform. However, it pays off in terms of academic achievement and is easily controllable by local officials [12]. Although it is said to be easily controlled by the government, the current state of the country's economy can be a major factor that will hinder the country's educational planning, including in resolving the issue of class size and lack of skilled teachers in programming. Something needs to be done so that with the existing workload, teachers still have time to prepare themselves with the skills needed for Computer Science subjects.

Other related issues are about, how are the responsible authorities going to ensure the standardization for the teaching of coding? If most developed countries use a school-based assessment system [13], assessment that focuses more on academic achievement through formal national examinations is still regard as the best practice so far for developing countries like Malaysia. While cultivating the knowledge without putting a burden to the student with examination, apart from general examinations such as PT3 and SPM, conducting one standard test to assess the level of mastery of programming skills among secondary schools' students has not yet been made uniformly. A good programming assessment system is able to form students who truly master the desired skills for the future, not just learn to pass the exam. To the best of the author's knowledge, no specific study has been conducted to evaluate the effectiveness of public examinations on the formation of programming skills.

### *3.3 Digital Divide*

The use of mobile technologies appears to be in line with the strategic goals in education besides facilitating and promoting learning anywhere and anytime [14]. Despite the complete and advance mobile infrastructure in the developed world, the digital divide still exists in developing countries [14]. As for Internet connection, even though Malaysia (together with Singapore and Brunei) has been listed as one of the three countries in South East Asia that have over 80% Internet penetration, there are certain places in Malaysia have poor Internet service. For example, in Kuala Lumpur, people are enjoying high-speed Internet up to 800 MBps. At the same time, in Sarawak (East Malaysia), the Internet speed is much slower, with some areas in the state without any access to Internet service [15]. Even when online access is available, some challenges persist. As a developing region, in Southeast Asia, many

students are from economically vulnerable families. Their access to computers is limited to school-provided computer labs, and many do not have access to unlimited Internet on their mobile devices [15]. The rural areas will have some issues in implementing Computer Science education in their schools due to administrative and facility barriers. To learn coding, the students are going to need not just the teachers but also the facilities. The schools that will have the facilities will be those from the wealthier and more developed regions. The urban areas will likely have the facilities and support, leaving the urban areas further behind. Despite of all the issues discussed before, digital divide is sometimes not seen as a problem because Computer Science subjects are only offered by schools that have computer labs and teachers trained in ICT. Or in other words, programming is just an optional skill or value-added for students rather than a compulsory subject.

### *3.4 Quality Digital Content*

In Malaysia, text books are used as the main teaching and learning resources. Undeniably, the quality of text books produced after the implementation of KSSM (Secondary School Standards-Based Curriculum) is very much improved with more exciting components such as activity, case study, project, animation and short information on innovation and daily applications of computer science. Apart from the appealing print layouts and illustrations, the success stories and achievements pertaining to computer science in Malaysia are also included as an element of motivation to the students. Most of the text books mentioned the careers in computer science. This is good indeed as even though Malaysia had been introduced to IT since the emergence of Multimedia Super Corridor (MSC) in year 1996, the careers in IT are still not widely known. Refer to [16–21] for all the electronic text books of Computer-science related subjects starting from year six to fifth-form. All the text books are published in Malaysia/Malay language. The contents of the books are all well-crafted with full of meaningful information. In the government policy, students are encouraged to learn at self-paced, do self-accessed and self-assessed [22]. However, the quality digital tool for coding subjects, developed tailor to the needs of national school students is yet to be created. Despite of the excellent quality of the current text books adopted by national schools' students, use of tools is really needed. A quality digital tool is believed to contribute to high quality lessons since they have potential to increase students' motivation, connect students to many information sources, support active in-class and out-class learning environments, and let instructors to allocate more time for facilitation [23]. By still using the case study conducted in Malaysia, there was a recommendation that wants application to be developed to solely focus on education, in which the students will not be able to access other things apart of the learning contents [14]. It can also be seen that the use of reiterative independent method has high potential to be adopted in future applications. Reiterative independent method, is a method that instills in students the skills

and mindset for learning new materials without being directly taught. This method is similar to the established method named Kumon [24, 25].

### 3.5 Assessments

There are 2 main examinations for national secondary schools, namely Sijil Pelajaran Malaysia (SPM) and PT3. The SPM or the Malaysian Certificate of Education, is a national examination taken by all fifth-form (17 years old students) secondary school students in Malaysia. Another one is PT3, which is a summative assessment to assess the academic achievement of students at the lower secondary level in Malaysia. PT3 is taken by all third-form (15 years old students) secondary school students in Malaysia. Table 1 shows the examinations conducted for computer science-related subjects offered starting from year 2019. Programming courses were assessed in these two major national examinations. As for school-level examination, there is no specific format imposed to the examinations' questions. The schools are free to construct the questions that suit to their school's students. The information of

**Table 1** The examinations conducted for computer science-related subjects offered starting from year 2019

| Year of study | Name of computer science-related subject | Examinations | Major examination format |
|---|---|---|---|
| Standard six<br>12 years old | Information and Communication Technology (ICT)<br>Note: Design Technology (optional if not taking ICT) | School-level examination | There is no major examination is conducted for this level of study for ICT subject. Only school-based exam is conducted |
| Form 1–3<br>13–15 years old | Basic Computer Science (BCS)<br>Note: Design Technology (optional if not taking BCS) | • School-level examination<br>• PT3 (in year 2020, PT3 examination was cancelled due to COVID-19 pandemic. No major examination announced as of date) | Year 2019 PT3 examination format:<br>• Written examination 70% (objectives and subjective questions)<br>• Project 30% |
| Form 4–5<br>16–17 years old | Computer Science (only available if the student takes Applied-Science package) | • School-level examination<br>• SPM | SPM 2020 examination format:<br>• Paper 1: written examination 70% (50 marks from close ended questions and 50 marks form open ended questions)<br>• Paper 2: project 30% |

major examinations in Malaysia are as of stated in [26]. The percentage of assessment for Basic Computer Science subject of PT3 examination is 70% allocated for written examinations and 30% for individual project that was assigned for the students to complete in about 6 months (from March to August). For the project, the PT3 candidates are required to translate ideas by writing step-by-step solutions in pseudocode and flow charts, which lastly translated into a program. The students are also required to analyze and compare method of program development, test, detect, and fix errors of program [27]. Example of project question for SPM candidates is to develop business information management system [28]. Since there is a need for students to master programming for the purpose of getting good grades for general examinations, this is seen to indirectly motivate students to practice what they have learned for the previous 2 years. Written examinations accompanied by projects are seen as complete enough to assess students' performance in mastering the subject. However, the question here is, is the learning process that students go through to face exams able to shape students towards mastering programming skills? Or, would it be possible that students who obtained A- and above in the major examination, actually do not have practical skills? One more thing, after completing SPM, students in Malaysia can choose whether to continue their studies by taking either a certificate, diploma, matriculation or sixth-form. At this stage if students choose the science stream, they will take subjects such as Mathematics, Physics, Biology, English and Chemistry. Only students who take the Foundation in Computer Science continue to study Computer Science related subjects such as basic programming, introduction to algorithms and some mathematics related subjects. It can be seen here that if students do not choose the basic course of Computer Science, there is a gap of one to two years before they continue their studies at the level of Computer Science degree. This results in what was previously learned cannot be practiced at the degree level.

It is recommended that assessments starting at the school level should lead to strengthening algorithm building and mastering syntax. For SPM level, written examinations for programming topics should be replaced with formatted questions allowing students to use compilers to get answers. The use of compilers can actually motivate students and make students so happy because they can see the results in front of their eyes. Project-based assessment is actually quite difficult for students under the age of 17, especially if they are still in the phase of exploring new fields. Projects can make students and teachers feel burdened.

### *3.6 Teaching and Learning Time*

Programming requires hands-on learning, at least 1–1.5 h a week. Students are usually able to improve their skills as they are gradually exposed to these new concepts gradually. Students also need a self-learning system to strengthen learning comprehension. However, since students need to study some other subjects, the time allotted for Computer-science subjects are very limited. The amount of time spent by the primary and secondary students in Malaysia to learn Computer Science is as shown in Table

**Table 2** The average teaching duration of computer science-related subjects offered starting from year 2019

| Year of study | Non-computer science subjects | | Computer science-related subject | |
|---|---|---|---|---|
| | List of subjects | Average teaching duration | List of subjects | Average teaching duration |
| Standard Six<br>12 years old | Malaysia Language, English, Mathematic, Islamic/Moral Study, Arabic Language, Health Study, Physical Education, Music, History, Visual Art Education | 6 h | Information and Communication Technology (ICT)<br>Note: Design Technology (an option if not take ICT) | 30 min |
| Form 1–3<br>13–15 years old | Malaysia Language, English, Mathematic, Islamic/Moral Study, Health Study, Physical Education, Music/Visual Art Education, History, Geography | 6.5 h | Basic Computer Science (BCS)<br>Note: Design Technology (an option if not take BCS) | 1 h |
| Form 4–5<br>16–17 years old | **Core subjects**<br>Malaysia Language, English, Mathematic, Islamic/Moral Study, History<br>**Compulsory subject**<br>Physical Education<br>**Elective subject**<br>Pure Science/Language/Islamic Study/Humanity/Applied Science-related subjects | 6.5 h | Science Computer (only available if the student takes Applied-Science package) | 1 h<br>30 min |

2. Despite of the limited time, we can see that the duration of teaching Computer Science increases as the level of study goes higher. Since finding solutions to the problem of limited learning time is quite difficult as it involves major issues such as logistics and financial allocation, the provision of a self-learning system needs to be provided. An in-house learning tool developed tailor to the specific needs of national schools' students and teachers is needed. Another issue that is worth to be discussed is pertaining to the interesting elements incorporated into the current text books. Direct link to the web site for additional information via the scanned QR codes and interesting project for the students to try, are some of the examples appealing current text book features of Malaysia schools. Undoubtedly, all these latest elements are very interesting and innovations like this are indeed to be commended. However, due to the time constraints, the students do not have much time to utilize all these new interesting features. Teachers have to work hard to finish the syllabus and students

struggle to also focus on other subjects. Again, a self-learning system needs to exist and it needs to continue to focus on practical skills.

## 4 Analysis

All issues presented in this paper are based on a study of the national education policy development plan as well as learning materials that are accessible to the general public. In the next study, we intend to obtain information directly from teachers, students and policy makers to verify validity. Based on the issues identified earlier, the main solution is short-term job exemption to give teachers the opportunity to create a simple system to practice programming.

Programming is not suitable for learning in situations where it is still necessary to perform other teaching tasks. In terms of academic assessment, all the solutions proposed in this paper are on the assumption that we still maintain the current syllabus which still requires students to learn about other theories, not programming alone. It is suggested that academic assessment should be based on a small percentage (30%) allocated to assess theory or knowledge in topics such data representation and algorithm reconstruction, and another 70% is for programming. Programming questions should not be in written format, instead allowing students to use a compiler to get answers. Examinations that take into account the project should be avoided because with the shortage of study time, it is feared that the project will be done not in a situation where students are interested in learning something but only to pass the exam. For long-term strategy, we should assess the sequence of the programming languages taught to the students. There are several issues that shall be considered which are presented in a form of the following questions; (1) is it necessary for the students to be exposed to several languages throughout their years of study? (2); did we consider the continuation of study? For example: After the student had completed learning Python in second-form and third-form, how does that knowledge is bringing forward to fourth-form and fifth-form?

## 5 Conclusions

The Malaysian Education System is undergoing a revolution where every year we can see there are so many improvements that have been made. Among the brilliant ideas that are being implemented is to introduce the subject of Computer Science to primary and secondary school students in Malaysia. Malaysia has invested heavily to develop coding skills among the students, with several plans have been formulated prudently in order to ensure the success of the mission. In pursuits of a mission, several issues and challenges have been identified with some possible solutions are presented. The identified issues are from the perspective of language of communication, implementation and execution, digital divide, quality tools, assessments and

teaching and learning time. Most of the recommended solutions to the identified issues do not require significant modifications to the existing education setting used by national schools in Malaysia. This works suggests creating a self-learning system built specifically for the national secondary schools' syllabus, short-term job exemption for teachers and programming skill test to replace project as part of assessments to increase the rate of teaching effectiveness of Computer Science subjects, especially programming.

**Acknowledgements** We would like to thank Universiti Tenaga Nasional for funding this study under the grant number RJ010517844/006.

## References

1. TheStar. https://www.thestar.com.my/news/education/2019/09/22/coding-in-national-schools. Last accessed 02 June 2020
2. Kanemune S, Shirai S, Tani S (2017) REPORTS informatics and programming education at primary and secondary schools in Japan. Olympiads Inf 11:143–150. https://doi.org/10.15388/ioi.2017.11
3. Tundjungsari V (2016) E-learning model for teaching programming language for secondary school students in Indonesia. In: Proceedings of 2016 13th international conference on remote engineering and virtual instrumentation, REV 2016, pp 262–266. https://doi.org/10.1109/REV.2016.7444477
4. Sklirou TS (2017) Programming in secondary education: applications, new trends and challenges. In: IEEE Global engineering education conference, EDUCON, pp 580–585. https://doi.org/10.1109/EDUCON.2017.7942904
5. Jawawi DNA, Mamat R, Ridzuan F, Khatibsyarbini M, Zaki MZM (2015) Introducing computer programming to secondary school students using mobile robots. In: 2015 10th Asian control conference: emerging control techniques for a sustainable world, ASCC 2015. https://doi.org/10.1109/ASCC.2015.7244750
6. Pang V. https://vulcanpost.com/582771/challenges-coding-education-malaysia/. Last accessed 06 June 2020
7. Ibrahim NM, Osman MM, Bachok S, Mohamed MZ (2016) Assessment on the condition of school facilities: case study of the selected public schools in Gombak district. Procedia Soc Behav Sci 222:228–234. https://doi.org/10.1016/j.sbspro.2016.05.151
8. Maros M, Kim Hua T, Salehuddin K (2007) Interference in learning English: grammatical errors in English essay writing among Rural Malay Secondary School students in Malaysia. e-BANGI Jurnal Sains Sosial dan Kemanusiaan 2(2):15
9. Asas Sains Komputer: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan. https://drive.google.com/file/d/11Mi25o8FKhQm6d4d4TNRQ15MWZ2HZaaH/view. Last accessed 30 June 2020
10. Olanrewaju A, Oluyomi K (2020) Students' interest and class size as predictive tools for academic achievement in physics. Int J Sci Res Publ 10(6):217. https://doi.org/10.29322/IJSRP.10.06.2020.p10227
11. Bringula RP, Aviles AD, Ymelda Batalla MC, Teresa Borebor MF, Anthony Uy MD, San Diego BE (2017) Modern education and computer science 5:1–8. https://doi.org/10.5815/ijmecs.2017.05.01
12. Mathis WJ (2017) The effectiveness of class size reduction psychosociological issues. Hum Resour Manage 5(1):176–183

13. Pentaksiran Berasaskan Sekolah PBS SPPBS. https://myschoolchildren.com/nSPPBS2.htm. Last accessed 05 Aug 2020
14. Sharina A, Latef A, Frohlich D, Calic J, Muhammad NH (2020). https://www.blueoceanstrategy.com/. Last accessed 03 July 2020
15. Jalli N (2020) Commentary: E-learning sees no smooth sailing in Malaysia and Indonesia. Channel News Asia, 07 Apr 2020
16. Buku Teks Teknologi Maklumat Komunikasi Tahun 6. https://anyflip.com/tqbf/zykl/. Last accessed 30 June 2020
17. Buku Teks Asas Sains Komputer 1. https://anyflip.com/dcnm/hpps. Last accessed 2020/06/30
18. Buku Teks Asas Sains Komputer Tingkatan 2. https://fliphtml5.com/cfdkq/snqq. Last accessed 30 June 2020
19. Buku Teks Sains Komputer Tingkatan 3. https://anyflip.com/usff/onyp/basic. Last accessed 30 June 2020
20. Buku Teks Sains Komputer Tingkatan 4. https://online.anyflip.com/wexi/bwqu/mobile/index.html. Last accessed 30 June 2020
21. Buku Teks Sains Komputer Tingkatan 5. https://online.anyflip.com/wexi/pjsx/mobile/index.html. Last accessed 30 June 2020
22. Standard D, Dan K, Tingkatan P, Kurikulum BP (2015) Sains Komputer: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 4
23. Cigdemoglu HAC (2016) Use of ICT tools and their effect on teaching and learning; students' and instructor's views. In: EDULEARN16 proceedings, pp 5318–5322
24. Agita A (2005) The effect of application Kumon learning method in learning mathematics of ability troubleshooting mathematics of students. J Phys Conf Ser 1429(1):1
25. Ukai N (1994) The Kumon approach to teaching and learning. J Jpn Stud 20(1):87–113
26. Malaysia Education Syndicate. https://lp.moe.gov.my/. Last accessed 30 June 2020
27. Panduan Kerja Projek ASK (Asas Sains Komputer) PT3 - Bumi Gemilang. https://www.bumigemilang.com/panduan-kerja-projek-ask-asas-sains-komputer-tingkatan-3-mulai-tahun-2019/. Last accessed 30 June 2020
28. Kerja Kursus Sains Komputer SPM 2020 (Tema). https://upuonline.net/kerja-kursus-sains-komputer/. Last accessed 30 June 2020
29. Dokumen Standard Kurikulum dan Pentaksiran Tahun 6 Teknologi Maklumat Komunikasi. https://www.moe.gov.my/muat-turun/penerbitan-dan-jurnal/dskp-kssr. Last accessed 30 June 2020

# The Similarity Finite Difference Solutions for Two-Dimensional Parabolic Partial Differential Equations via SOR Iteration


**N. A. M. Ali, J. Sulaiman, A. Saudi, and N. S. Mohamad**

**Abstract** This paper purposely attempts to solve two-dimensional (2D) parabolic partial differential equations (PDEs) using iterative numerical technique. Also, we determine the capability of proposed iterative technique known as Successive Over-Relaxation (SOR) iteration compared to Gauss–Seidel (GS) iteration for solving the 2D parabolic PDEs problem. Firstly, we transform the 2D parabolic PDEs into 2D elliptic PDEs then discretize it using the similarity finite difference (SFD) scheme in order to construct a SFD approximation equation. Then, the SFD approximation equation yields a large-scale and sparse linear system. Next, the linear system is solved by using the proposed iterative numerical technique as described before. Furthermore, the formulation and implementation of SOR iteration are also included. In addition to that, three numerical experiments were carried out to verify the performance of the SOR iteration. Finally, the findings show that the SOR iteration performs better than the GS iteration with less iteration number and computational time.

**Keywords** SOR iteration · Similarity finite difference scheme · Two-dimensional parabolic partial differential equations

## 1 Introduction

Similarity solution is a famous technique in transform partial differential equations (PDEs) to ordinary differential equations (ODEs) which has been applied by many

N. A. M. Ali (✉) · J. Sulaiman · N. S. Mohamad
Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
e-mail: afzamatali@yahoo.com

J. Sulaiman
e-mail: jumat@ums.edu.my

N. S. Mohamad
e-mail: norsyahida1302@gmail.com

A. Saudi
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
e-mail: azali@ums.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_42

researchers to solve different problems, see [1–6]. However, different approach compared to other researchers, in this paper, we use the similarity transformation via wave variables to transform 2D parabolic PDEs into 2D elliptic PDEs. This is because we want to remain the characteristic of its solution domain without time.

Based on the previous study, there are many researchers have solved their generated linear system by using the most known and widely used iterative technique namely Successive Over-Relaxation (SOR) iteration [7–12] in which was proposed by Young [13]. Nevertheless, there is no study that has been conducted in literature for solving 2D parabolic PDEs using SOR iteration via the similarity finite difference (SFD) scheme. Therefore, with the superiority of the SOR iteration in term of its computational cost and similarity solutions, we make an effort to determine the capability of SOR iteration in order to solve the 2D parabolic PDEs as compared to Gauss–Seidel (GS) iteration.

The capability of SOR iteration is determined by imposing the 2D parabolic PDEs stated as

$$\frac{\partial u}{\partial t} + \beta_1 \frac{\partial u}{\partial x} + \beta_2 \frac{\partial u}{\partial y} = \alpha_1 \frac{\partial^2 u}{\partial^2 x} + \alpha_2 \frac{\partial^2 u}{\partial^2 y} + R(x, y, t), \tag{1}$$

subject to the initial condition

$$u(x, y, a) = f(x, y),$$

and boundary condition

$$u(a, y, t) = g_1(y, t), u(b, y, t) = g_2(y, t),$$

$$u(x, a, t) = h_1(x, t), u(x, b, t) = h_2(x, t),$$

with the solution domain, $D = [a, b] \times [a, b] \times [a, T]$.

As mentioned in previous studies [14, 15], all researchers have transformed 2D problems into one-dimensional (1D) two boundary value problems via the wave variables transforms as follows

$$\xi = x - \phi_0 t, \xi = y - \phi_1 t \tag{2}$$

where $\phi_0$ and $\phi_1$ are constants. In order to remain the characteristic of the solution domain of problem (1) without time, we initiate the following wave variables transformations

$$\xi = x - ct, \tau = y - dt \tag{3}$$

where $c$ and $d$ are constants. The problem (1) is reduced to 2D elliptic PDEs by applying (3) into (1) becomes

$$\alpha_1 \frac{d^2u}{d\xi^2} + \alpha_2 \frac{d^2u}{d\tau^2} + \left(\frac{c}{2} - \beta_1\right)\frac{du}{d\xi} + \left(\frac{d}{2} - \beta_2\right)\frac{du}{d\tau} = -R(\xi, \tau) \tag{4}$$

The rest of this paper is structured as follows. In Sect. 2, we discussed the derivation of SFD approximation equations based on SFD schemes. The approach of proposed iterative technique will be explained in Sect. 3. Next, some numerical results were presented in Sect. 4 to show the capability of the proposed method. We conclude the results in final section.

## 2 Derivation of Similarity Finite Difference Approximation Equation

Before discretizing the problem (4), we need to built the finite grid network in which this network is an important part to guide us in development and implementation of proposed iterative numerical technique. Thus, we show the finite grid network at $n = 7$ as depicted in Fig. 1.

According to Fig. 1, we need to discretize uniformly the solution domain, $(\xi, \tau)$ in both $\xi$ and $\tau$ directions with a mesh size, $h$ which is defined as

$$h = \frac{\tau - \xi}{m}, m = n + 1 \tag{5}$$

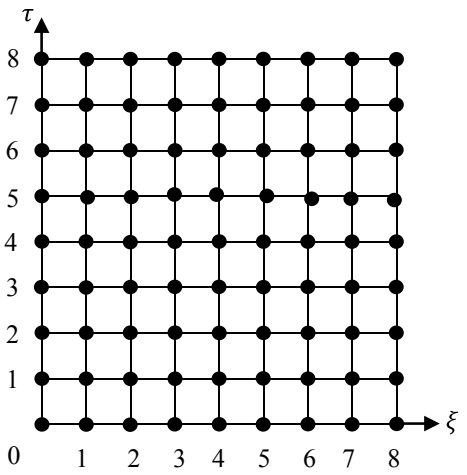



**Fig. 1** Finite grid networks at $n = 7$

Having the finite grid network in Fig. 1, we discretize the 2D elliptic PDEs (4) by using the SFD scheme as follows:

$$\left.\begin{aligned} \frac{du}{d\xi}\Big|_{ij} &= \frac{U_{i+1,j}-U_{i-1,j}}{2h}, \\ \frac{d^2u}{d\xi^2}\Big|_{ij} &= \frac{U_{i+1,j}-2U_{i,j}+U_{i-1,j}}{h^2}, \\ \frac{du}{d\tau}\Big|_{ij} &= \frac{U_{i,j+1}-U_{i,j-1}}{2h}, \\ \frac{d^2u}{d\tau^2}\Big|_{ij} &= \frac{U_{i,j+1}-2U_{i,j}+U_{i,j-1}}{h^2}. \end{aligned}\right\} \tag{6}$$

Then, we substitute (6) into (4) and multiply by $h^2$ yields

$$\alpha_1\big(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}\big) + \alpha_2\big(U_{i,j+1} - 2U_{i,j} + U_{i,j-1}\big)$$

$$+\big(U_{i+1,j} - U_{i-1,j}\big) + \gamma_1\big(U_{i,j+1} - U_{i,j-1}\big) = -h^2 R_{i,j} \tag{7}$$

where

$$\gamma_0 = h\left(\frac{c - 2\beta_1}{4}\right),$$

$$\gamma_1 = h\left(\frac{d - 2\beta_2}{4}\right).$$

To reduce the computational complexity, the SFD approximation Eq. (7) is simplified to

$$AU_{i-1,j} + BU_{i+1,j} + CU_{i,j-1} + DU_{i,j+1} + EU_{i,j} = -h^2 R_{i,j} \tag{8}$$

where

$$A = \alpha_1 - \gamma_0,$$

$$B = \alpha_1 + \gamma_0,$$

$$C = \alpha_2 - \gamma_1,$$

$$D = \alpha_2 + \gamma_1,$$

$$E = -2\alpha_1 - 2\alpha_2.$$

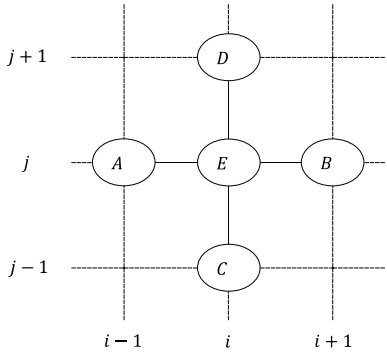



**Fig. 2** Computational molecules of the similarity approximation equation

Actually, the approximation Eq. (8) can be diagrammatically represented by its computational molecules given in Fig. 2.

Also, the SFD approximation Eq. (8) can be used to construct a similarity linear system and may be written in matrix form as

$$F\underline{U} = \underline{R} \tag{9}$$

where

$$F = \begin{bmatrix} G_2 & G_3 & 0 & 0 & \cdots & 0 \\ G_1 & G_2 & G_3 & 0 & \cdots & 0 \\ 0 & G_1 & G_2 & G_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & G_1 & G_2 & G_3 \\ 0 & 0 & 0 & 0 & G_1 & G_2 \end{bmatrix}, \underline{U} = \begin{bmatrix} \underline{U}_1 \\ \underline{U}_2 \\ \underline{U}_3 \\ \underline{U}_4 \\ \vdots \\ \underline{U}_n \end{bmatrix}, \underline{R} = \begin{bmatrix} \underline{R}_1 \\ \underline{R}_2 \\ \underline{R}_3 \\ \underline{R}_4 \\ \vdots \\ \underline{R}_n \end{bmatrix}$$

and

$$G_1 = \begin{bmatrix} C & 0 & 0 & \cdots & 0 \\ 0 & C & 0 & \cdots & 0 \\ 0 & 0 & C & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & C \end{bmatrix}, G_2 = \begin{bmatrix} E & B & 0 & \cdots & 0 \\ A & E & B & \cdots & 0 \\ 0 & A & E & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & A & E \end{bmatrix}, G_3 = \begin{bmatrix} D & 0 & 0 & \cdots & 0 \\ 0 & D & 0 & \cdots & 0 \\ 0 & 0 & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & D \end{bmatrix},$$

$\underline{U}_i = \begin{bmatrix} U_{1,i} & U_{2,i} & U_{3,i} & U_{4,i} & \cdots & U_{n,i} \end{bmatrix}^T$,

$\underline{R}_i = \begin{bmatrix} -h^2 R_{1,i} & -h^2 R_{2,i} & -h^2 R_{3,i} & -h^2 R_{4,i} & \cdots & -h^2 R_{n,i} \end{bmatrix}^T$,

with $i = 1, 2, 3, \ldots, m-1$.

## 3 SOR Iteration Approach

In an effort to formulate the SOR iteration and solve the linear system (9), let the coefficient of matrix $F$ in Eq. (9) be decomposed as summation of three matrices as

$$F = D - L - V \tag{10}$$

where $D$, $L$ and $V$ are diagonal, lower triangular and upper triangular matrices. From the decomposition (10), the SOR iteration can be stated generally as follows: see, [13, 16, 17]

$$\underline{U}^{(k+1)} = (D - \omega L)^{-1}\left[(\omega V - (1 - \omega L)D)\underline{U}^{(k)}\right] + (D - \omega L)^{-1}\underline{R} \tag{11}$$

where $\omega$ indicates a relaxation factor. Thus, the implementation of the SOR iteration would be described in Algorithm 1.

Algoritm 1:

i. Initialize $\underline{u}^{(0)}$ and $\varepsilon = 10^{-10}$
ii. Calculate coefficient matrix, $F$ and vector, $\underline{R}$
iii. Calculate,
$$\underline{U}^{(k+1)} = (D - \omega L)^{-1}\left[(\omega V - (1 - \omega L)D)\underline{U}^{(k)}\right] + (D - \omega L)^{-1}\underline{R}$$
iv. Check the convergence test, $\left|\underline{U}^{(k+1)} - \underline{U}^{(k)}\right| \leq \varepsilon = 10^{-10}$. If satisfied, go to step (v). Otherwise go back to step (iii).
v. Display approximate solutions.

## 4 Numerical Experiments

In this section, we make a comparative analysis between the SOR and GS iterations to determine the capability of SOR iteration by exemplifying three selected problems of 2D parabolic PDEs. In comparison for these iterative methods, three measurement parameters are considered such as iteration number ($k$), computational time in second ($t$) and maximum absolute error ($Err$). Also, five different mesh sizes ($m$) which are 64, 128, 256, 512 and 1024 are considered in these numerical experiments. Among the three 2D parabolic PDEs problems are as follows:

**Problem 1:** See in [18]

$$\frac{\partial u}{\partial t} + \beta_1 \frac{\partial u}{\partial x} + \beta_2 \frac{\partial u}{\partial y} = \alpha_1 \frac{\partial^2 u}{\partial x^2} + \alpha_1 \frac{\partial^2 u}{\partial y^2}, \tag{12}$$

exact solution

$$u(x, y, t) = \frac{1}{4t+1} \exp\left(-\frac{(x-0.8t-0.5)^2}{0.01(4t+1)} - \frac{(y-0.8t-0.5)^2}{0.01(4t+1)}\right) \tag{13}$$

**Problem 2:** See in [19]

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \sin x \sin y e^{-t} - 4, \tag{14}$$

exact solution.

$$u(x, y, t) = \sin x \sin y e^{-t} + x^2 + y^2. \tag{15}$$

**Problem 3:** See in [20]

$$\frac{\partial U}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \tag{16}$$

exact solution.

$$u(x, y, t) = e^{-5\pi^2 t} \sin(2\pi x) \sin(\pi y). \tag{17}$$

A thorough comparison of the numerical results is presented clearly in Tables 1, 2 and 3, as well as in Figs. 3, 4, 5, 6, 7 and 8 in order to determine the capability of SOR iteration compared to GS iteration. In addition, we also present in Table 4 about the percentage of reductions for SOR iteration compared to GS iteration in terms of iteration number and computational time over proposed problems.

Based on the all numerical results tabulated in Tables 1, 2 and 3, it can be pointed out that the SOR iteration requires less iteration number compared to GS iteration which can be depicted in Figs. 3, 4 and 5. Also, we can see that SOR iteration needs

**Table 1** Numerical results of Problem 1

| *m* | Method | *k* | *t* | *Err* |
|---|---|---|---|---|
| 64 | GS | 91 | 0.07 | 4.540422E−04 |
| | SOR | 12 | 0.00 | 4.540422E−04 |
| 128 | GS | 494 | 0.39 | 4.543391E−04 |
| | SOR | 25 | 0.04 | 4.543391E−04 |
| 256 | GS | 2152 | 5.61 | 4.544134E−04 |
| | SOR | 50 | 0.16 | 4.544134E−04 |
| 512 | GS | 8742 | 91.23 | 4.544325E−04 |
| | SOR | 98 | 1.12 | 4.544325E−04 |
| 1024 | GS | 34,380 | 1434.90 | 4.544364E−04 |
| | SOR | 198 | 9.99 | 4.544373E−04 |

**Table 2** Numerical result of Problem 2

| *m* | Method | *k* | *t* | *Err* |
|---|---|---|---|---|
| 64 | GS | 7104 | 1.67 | 5.884082E−03 |
| | SOR | 266 | 0.10 | 5.884048E−03 |
| 128 | GS | 26,089 | 15.78 | 5.885178E−03 |
| | SOR | 525 | 0.41 | 5.885038E−03 |
| 256 | GS | 95,090 | 209.01 | 5.886575E−03 |
| | SOR | 1036 | 2.57 | 5.886017E−03 |
| 512 | GS | 343,412 | 3023.04 | 5.888316E−03 |
| | SOR | 2060 | 20.13 | 5.886079E−03 |
| 1024 | GS | 1,226,108 | 43,176.41 | 5.895053E−03 |
| | SOR | 4108 | 163.14 | 5.886095E−03 |

**Table 3** Numerical result of Problem 3

| *m* | Method | *k* | *t* | *Err* |
|---|---|---|---|---|
| 64 | GS | 2067 | 0.60 | 7.171921E−03 |
| | SOR | 138 | 0.05 | 7.171949E−03 |
| 128 | GS | 5930 | 3.62 | 7.171831E−03 |
| | SOR | 267 | 0.21 | 7.171948E−03 |
| 256 | GS | 14,541 | 32.92 | 7.171463E−03 |
| | SOR | 524 | 1.35 | 7.171947E−03 |
| 512 | GS | 27,980 | 249.82 | 7.170091E−03 |
| | SOR | 1035 | 11.18 | 7.171947E−03 |
| 1024 | GS | 38,776 | 1429.23 | 7.166526E−03 |
| | SOR | 2054 | 80.34 | 7.171940E−03 |

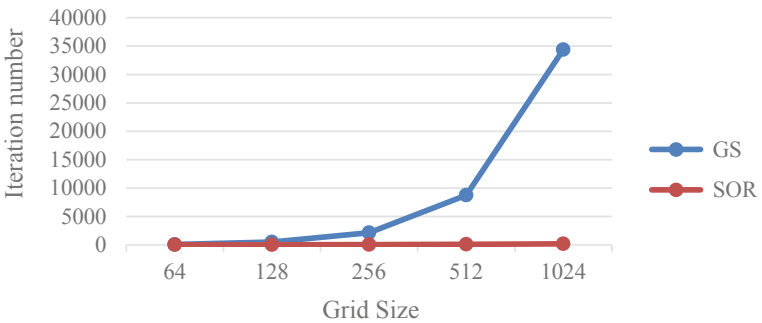



**Fig. 3** Comparison of iteration number for SOR and GS iteration over Problem 1

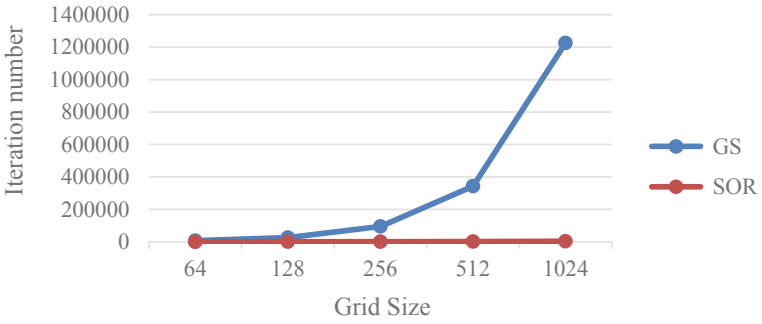



**Fig. 4** Comparison of iteration number for SOR and GS iteration over Problem 2

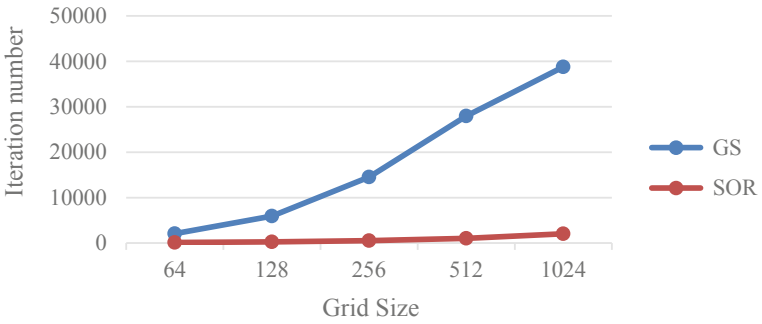



**Fig. 5** Comparison of iteration number for SOR and GS iteration over Problem 3

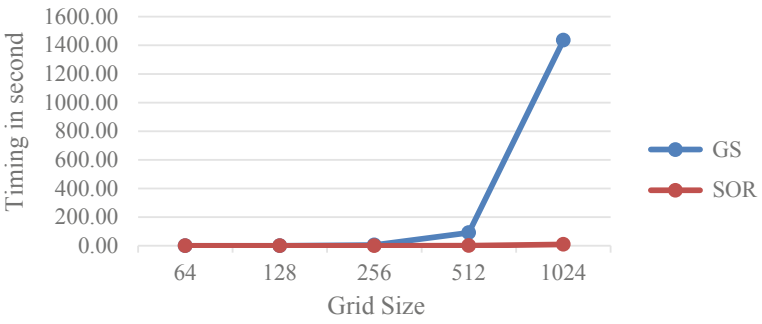



**Fig. 6** Comparison of computational time for SOR and GS iteration over Problem 1

a shorter computational time compared to GS iteration in Tables 1, 2 and 3 and Figs. 6, 7 and 8. Moreover, Table 4 shows the percentage of decrement in terms of iteration number have declined approximately by 86.81–99.42%, 96.26–99.67% and 93.32–96.40% respectively correspond to SOR iteration compared to GS iteration.

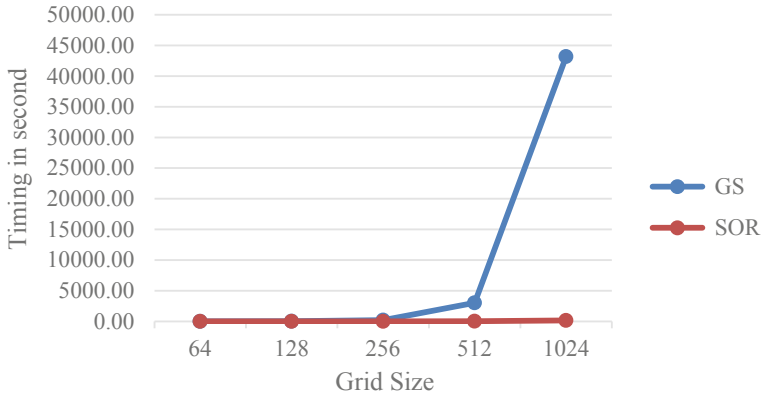



**Fig. 7** Comparison of computational time for SOR and GS iteration over Problem 2

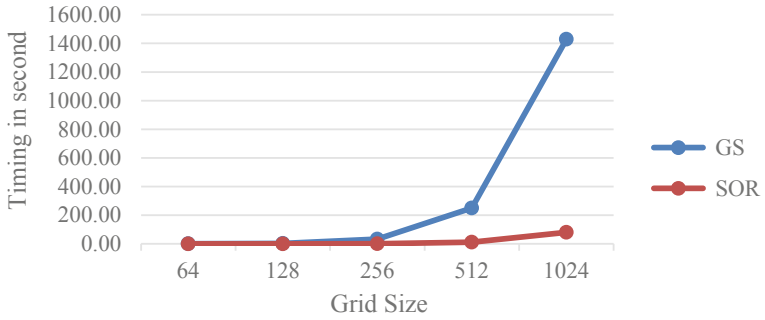



**Fig. 8** Comparison of computational time for SOR and GS iteration over Problem 3

**Table 4** Decrement percentage of iteration number and computational time for SOR iteration compared to GS iteration over proposed problems

| Problem | Iteration number (%) | Computational time (%) |
|---|---|---|
| 1 | 86.81–99.42 | 89.74–99.00 |
| 2 | 96.26–99.67 | 94.01–99.62 |
| 3 | 93.32–96.40 | 91.67–95.90 |

Meanwhile, the implementations of SOR iteration are much faster than GS iteration about 89.74–99.99%, 94.01–99.62% and 91.67–95.90% respectively.

## 5 Conclusions

In this paper, we have proposed SOR iteration to solve similarity linear system arising from the discretization of 2D elliptic PDEs problems using similarity finite difference

scheme. Obviously, it can be concluded that the numerical solution obtained by the SOR iteration is superior than GS iteration with smaller iteration number and shorter computational time. For future works, we extended our study by using half-sweep iteration concept [21–23] to solve 2D parabolic PDEs problems.

**Acknowledgements** The authors gratefully acknowledge the financial support from the Postgraduate Centre Universiti Malaysia Sabah for this research work.

## References

1. Makinde OD (2010) Similarity solution of hydromagnetic heat and mass transfer over a vertical plate with a convective surface boundary condition. Int J Phys Sci 5(6):700–710
2. Kolomenskiy D, Moffatt HK (2012) Similarity solutions for unsteady stagnation point flow. J Fluid Mech 711:394–410
3. Zhou H, Kong G, Liu H, Laloui L (2017) Similarity solution for cavity expansion in thermoplastic soil. Int J Numer Anal Methods Geomech 42(2):274–294
4. Paliathanasis A (2019) Similarity solutions for the wheeler-deWitt equation in $f(R)$-cosmology. Eur Phys J C 79:1031
5. Khashi'ie NS, Arifin NM, Pop I, Nazar R, Hafidzuddin EH (2020) A new similarity solution with stability analysis for the three-dimensional boundary layer of hybrid nanofluids. Int J Numer Methods Heat Fluid Flow
6. Khuri SA (2007) Similarity solution of the mixed convection boundary-layer flow in a porous medium. Int J Comput Methods 4(4):621–631
7. Rahman R, Ali NAM, Sulaiman J, Muhiddin FA (2018) Caputo's finite difference solution of fractional two-point boundary value problems using SOR iteration. In: International conference on mathematics, engineering and industrial applications. AIP Publishing, Malaysia, p 020034
8. Justine H, Sulaiman J (2017) Solution of fourth-order two-point BVPs with cubic non-polynomial spline and SOR iterative method. J Fundam Appl Sci 9(5S):579–593
9. Sunarto A, Sulaiman J, Saudi A (2014) Full-sweep SOR iterative method to solve space-fractional diffusion equations. Aust J Basic Appl Sci 8(24):153–158
10. Zhang C, Xue Z, Luo S (2016) A convergence analysis of SOR iterative methods for linear systems with weak $H$-matrices. Open Math 14:747–760
11. Mai T, Wu L (2013) The successive over relaxation method in multi-layer grid refinement scheme. Adv Appl Sci Res 4(2):163–168
12. Rivaz A, Abad FSPS (2014) Gauss-seidel and successive over relaxation iterative methods for solving system of fuzzy sylverter equations. J Mahani Math Res Cent 3(1):51–60
13. Young DM (1954) Iterative method for solving partial differential equation elliptic type. Trans Am Math Soc 76:92–111
14. Ali NAM, Rahman R, Sulaiman J, Ghazali K (2018) SOR iterative method with wave variable transformation for solving advection-diffusion equations. In: International conference on mathematics, engineering and industrial applications. AIP Publishing, Malaysia, p 020036
15. Bibi S, Mohyud-Din ST (2014) Travelling wave solutions of KdVs using sine-cosine method. J Assoc Arab Univ Basic Appl Sci 15:90–93
16. Hackbusch W (1995) Iterative solution of large sparse systems of equations. Springer, New York
17. Alibubin MU, Sunarto A, Akhir MKM, Sulaiman J (2016) Performance analysis of half-sweep SOR iteration with rotated nonlocal arithmetic mean scheme for 2D nonlinear elliptic problems. Glob J Pure Appl Math 12(4):3415–3424
18. Shallal MAM, Jumaa BF (2016) Numerical solutions based on finite difference techniques for two-dimensional advection-diffusion equation. Brit J Math Comput Sci 16(2):1–11

19. Ibrahim A, Abdullah AR (1995) Solving the two-dimensional diffusion equation by the four-point explicit decoupled group (EDG) iterative method. Int J Comput Math 58(3–4):253–263
20. Evans DJ, Sahimi MS (1988) The alternating group explicit (AGE) iterative method for solving parabolic equations 1: 2-dimensional problems. Int J Comput Math 24(3–4):311–341
21. Ali NAM, Rahman R, Sulaiman J, Ghazali K (2019) Solutions of reaction-diffusion equations using similarity reduction and HSSOR iteration. Indonesian J Electr Eng Comput Sci 16(3):1430–1438
22. Rahman R, Ali NAM, Sulaiman J, Muhiddin FA (2019) Application of the half-sweep EGSOR iteration for two-point boundary value problems of fractional order. Adv Sci Technol Eng Syst J 4(2):237–243
23. Zainal NFA, Sulaiman J, Alibubin MU (2019) Application of half-sweep iteration with nonlocal arithmetic discretization scheme for solving Burger's equation. ARPN J Eng Appl Sci 14:616–621

# JKalvi: An E-Learning Game Approach


**Darveen Selvarajah, Vinesha Selvarajah, and Ji-Jian Chin**

**Abstract** This paper discusses the implementation of "JKalvi", an E-Learning approach used to replace traditional teaching and learning methods through the use of E-Games. Although the use of educational technology in education is leveraging, there are still gaps in the use and implementation of such approaches being utilized for teaching and learning especially in high school education. In this study, users, both educators and students can utilize the resources in the application in increasing the interactivity of learning through the use of online materials and games all in a simple android platform. Given the increased use of android based mobile phones among high school students today, users can monitor signs of progress and performance through a digitized system replacing traditional paper-based methods all from a click away through their mobile phone. This is persistent with the examination of students' perceptions in the pursuit of new learning methodology through the use of this application conducted through via surveys. A total of 30 preliminary surveys were recorded and analyzed using descriptive statistics. The findings indicate that 90% of the students agreed that the application was interactive, added value to their learning process and found no issues in the use of the application.

**Keywords** Educational technology · High school students · E-Learning · Game-based · Application

D. Selvarajah (✉)
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
e-mail: darveenselvarajah@gmail.com

V. Selvarajah
Faculty of Computing, Engineering, and Technology, Asia Pacific University, Technology Park Malaysia, Kuala Lumpur, Malaysia
e-mail: vinesha.selvarajah@monash.edu

J.-J. Chin
Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
e-mail: jjchin@mmu.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_43

# 1 Introduction

The integration of the E-Learning platform as an educational technology goes back to its first discovery in 1924 [9, 13]. Numerous research had revolved in the area of increasing the effectiveness of E-Learning application in promoting steep learning curve especially among high school students [11]. In the recent years, the pursuit of integrating technology into the education system has become increasingly popular due to the vast advancement of technology [2]. This has led to the birth of the term "Educational Technology" especially in the area of elementary, secondary and university students [1, 14].

The future of education is educational technology. This enables learning to be seamlessly accessible from anywhere and anytime through the online platform [5]. This increases the efficiency in managing content by the educator and student without just the face-to-face mode of learning. The progression of the implementation of education technology then allows for queries and analytics where data has been digitized. With proper algorithms and analytics, educators and students can assess their progress and the competency of student performance instead of a traditional paper-based method. This helps to save time and cost consumptions in leveraging the use of technology in education.

## *1.1 Types of Educational Technology*

Through the years of research and development in education technology, various types of educational technology are in place [10]. Such examples are: flipped classroom learning, electronic whiteboards, desktops and laptops, video conferencing classroom technologies such as Skype, Google Hangouts and even Microsoft Teams [15]. Mobile learning and televisions educational learning has also impacted the way education is catered. This enables the monitoring of student progress and achievement through a central point. This gives the educator more control on how to manage the teaching and learning of knowledge in the classroom. This also promotes the facilitation of distant learning for students and educators who have issues in access to resources in rural areas [12]. This allows virtual travel for both educators and students in teaching and learning including modules such as space knowledge without physically travelling to space [6].

## *1.2 Game-Based E-Learning*

Game-based E-Learning has various fields to assist students in learning. For example, with the current implementation of education, many students find it difficult to adapt

to the learning process. Many students find it more intriguing in learning with Game-based E-Learning rather than the traditional approach [9]. The main idea of this paper is to deploy an E-Learning system rapidly with less effort and more compromising quality of the learning practice. The game-based application is a strategy that is developed to enhance learner's motivation in gaining new knowledge. For example, with colourful image and effects and diversified teaching materials, this can be used to enhance the student's learning motivation and further improvise the effectiveness of learning.

## 2 Proposed Application

The proposed application was designed to replace traditional based learning in cultivating the leveraged use of technology in promoting efficient education. The sections below describe the differences between the approach used in this study and the traditional learning approach.

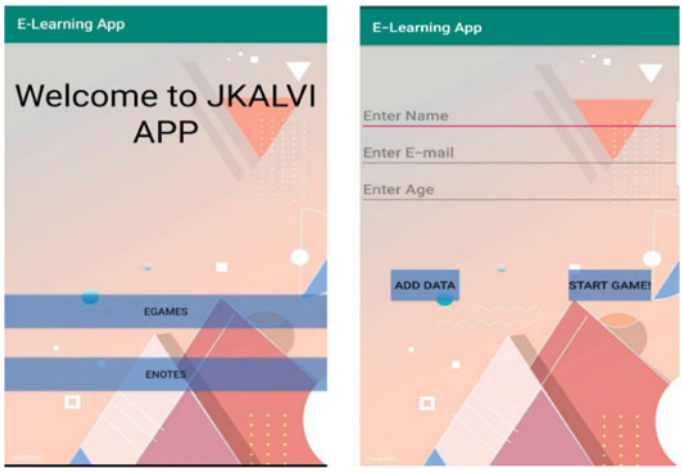
### *2.1 Traditional Learning Versus E-Learning*

Traditional learning is an ancient method of learning that is conducted by a teacher gathering students in places such as classes, labs or seminars to study and learn about different subjects. This method of learning has been practiced around the world in all levels such as kindergartens, primary, secondary, high schools, colleges and universities. The students will be able to ask questions directly to the teacher and exchange information immediately. The traditional approach is more suitable for the younger children and students for primary as they are not very tech-savvy yet [7].

E-Learning and the traditional approach vary a lot; E-Learning is a virtual class that can happen anywhere with the need for a smartphone, laptop and a connection. The application created is basically to help students from urban areas that have transport and difficulty in moving to the city area. E-Learning is not only based on a software application, but it has the possibility of having an online interactive class, and video conferencing class.

Based on the proposed application that was built, the game was intended to assess knowledge from mathematical theories. Each game will start with a section of registration of the student's info. There are two sections of the application: The first section is the E-Notes where by students can access the structure in learning mathematics. The second part consists of the part where students will then choose the option to move on to the game part; where the application provides the students with questions such as MCQ, true and false, and structured questions.

The process was developed from as early as the '90s. The whole idea was to ease the burden of the traditional approach where it is costly to come out to gain knowledge




(a) Main Page (b) Registration Page

**Fig. 1** The above images illustrate the **a** main page of the E-Learning application prototype called "JKalvi App". The application allows users to select whether there are interested in going through the E-Learning Notes or E-Games related to the subject area. Image in **b** illustrates the registration page where users can register themselves into the systems to retrieve relatable content based on their age and education level

[4]. Along with previous development in E-Learning platforms, an interactive game-based challenge was used in this research study in educating and assessing the skills in mathematics.

## 3 Design and Implementation

This section describes the design and implementation of the JKalvi E-Learning Game-Based application in this study. This section first explains the design of each page in the application. The main page is illustrated in Fig. 1.

## 4 Methodology Used

The software development methodology used in developing the JKalvi Application is the waterfall model [3]. The main reason the waterfall model is used is to ensure that each component is completed before moving to other sections of the application. The waterfall method is also one of the easiest methods to implement in building the android application [8]. The steps that were involved and integrated into the development of this application include (I) Requirement and analysis, (II) System

**Table 1** The use of the waterfall model in the development of the E-Learning Game-Based application

| | |
|---|---|
| Requirement and analysis | This phase explains the objectives and why do students and why do students slack in the use of E-Learning application. We collect the data from users that are to be implemented later on |
| System design | The system is built by making sure that the interface is user friendly and will be easy to be used and accessible by students |
| System implementation | After the verification and sufficient collection of information on how the system would be, the system is then coded to be implemented |
| System testing | The system is tested by the team to check for errors and bugs. The survey given to students is also part of the system testing |
| Deployment of the system | The software is ready to be used once the testing phase is done |
| System maintenance | This is where the regular updating, verification and debugging on the system in making sure the application runs smoothly |

Design, (III) System Implementation, (IV) System Testing, (V) Deployment of the system and (VI) System maintenance.

The steps undertaken in the waterfall model in this study is further described in Table 1:

The flow of the application is illustrated in the following flowchart under Fig. 2.

Subsequently, a survey was administered after the development of the application to elicit the feedback from participants, who were high school students. The survey consists of 4 questions in which each participant had to provide their feedback in the perception of the use of the application. The following questions were asked during the survey, with feedback presented in the next section:

1. How do you find the overall system?
2. How do you find the functionality of the system?
3. What are the challenges you faced while using the application?
4. Would you recommend the use of this app to your peers?

## 5 Results and Analysis

A total of 30 surveys were collected during the data collection process. The students were briefed on the use of the application before downloading the app to their android based mobile phone. The demographic details of the participant are shown in Table 2.

The participants had downloaded the application using the following spectrum of mobile phones, Samsung Galaxy Note 5 and OPPO F11 Pro both which operates on Android OS. Approximately 90% ($f = 27$) of the total participants have indicated the use of the application to be good, and user friendly. In terms of functionality of the system, approximately 70% of the participants reported not having any issues,

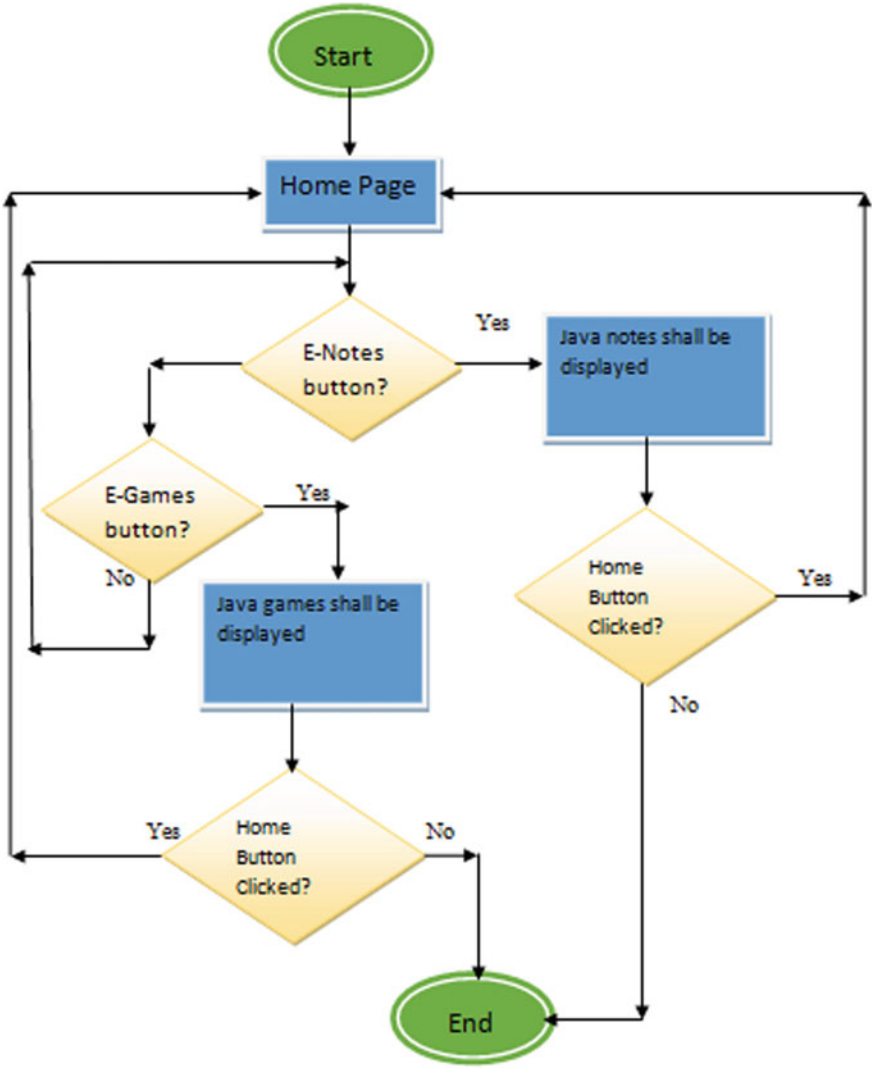



**Fig. 2** JKalvi application flow chart

**Table 2** Demographic details of participants

| Demographics | | Count (%) |
|---|---|---|
| Gender | Female | 12 (40%) |
| | Male | 18 (60%) |
| Age group | 14 years | 19 (63%) |
| | 15 years | 11 (37%) |
| Time taken to complete testing | <15 min | 26 (85%) |
| | >15 min | 4 (15%) |

while the remaining suggested having more functionality or subjects to be integrated. However, since this an only a prototype implementation, further functionalities will be included under future works. On the other, as for the challenges and recommendation, all participants have indicated no challenges faced throughout the application and would recommend the use of this app to their peers once more functionality has been integrated.

The research and analysis done were based on the application that is in the Google play store. The applications that were taken into comparison with JKalvi are (i) Delegate E-Learning, (ii) IILT's E-Learning platform, and (iii) MCIS Life eLearning applications. All the three application is implemented with such functionality that users have to register their credentials to use the application. Delegate E-Learning requires users to learn the functions of the QR Code scanner to use the application. These applications also require the user to purchase the application to use all the functionality of the application. JKalvi is a more user-friendly application as the interface is simple for users to use. To use the application, users are just required to key in their details such as Name, Age and email address. The main purpose to key in these details is for the users to obtain their score in the game based part. Users will automatically receive the marks through email. JKalvi navigation page is very direct and easy to access as the user can use all the functionality of the application upon putting in their details. The application does not require users to register.

Delegate application requires users to register an account to use the application, the application also uses the functionality of QR code scanner that requires the user to understand the concept of QR Code reader and how it operates. IILT's E-Learning platform requires the user to register their credentials before using the application. CIS Life eLearning application requires the user to register and login using their credentials.

JKalvi game-based application can be used by (i) students, and (ii) educators. Students will be able to obtain E-Notes from the application and use as references. Students will also be able to test their skills upon going through the E-Notes to see how effective the E-Notes are by solving questions in the E-Games that are related to the E-Notes in the application.

JKalvi provide educators with the flexibility of uploading E-Notes. They will also be able to keep track and monitor the progress of their student in JKalvi. Educators will have the options to test their students' skills or progress by uploading quizzes and tasks in the E-Game section of the application.

## 6 Discussion and Future Work

This research paper explored the development of the Jkalvi App, an E-Learning game-based application aiming at high school students in promoting knowledge and learning related to one of their modules learn in school—Mathematics. The application was developed to be used on an android platform for students to seek understanding of subject matters through the use of E-Notes and gaming. A survey was

administered right after the testing of the application with a total of 30 participants who participated in the application testing. Overall, the participants were satisfied with the JKalvi app and would look forward to using and recommending the application to their peers once more functionalities have been added. As the application in this study was designed to be used as a prototype, the only mathematic module was included.

The future work of this research study takes into consideration the recommendations given by the participants. Firstly, added features including several interactive games concerning other areas among the high school syllabus will be included. Furthermore, the design of the application would be further improved to increase the interactivity and aesthetics of the applications since the users of the system revolve around high school students. Lastly, the future focus would centre on developing the application which would be able to run across different platforms such as desktop versions and mobile version taking into the consideration of Macintosh operating systems as well.

**Acknowledgements** The authors would like to thank the Ministry of Education of Malaysia in providing financial support for this work through the Fundamental Research Grant Scheme (FRGS/1/2019/ICT04/MMU/02/5).

## References

1. Apple MW, Bromley H (1998) Education, technology, power: educational computing as a social practice. Suny Press
2. Beldarrain Y (2006) Distance education trends: integrating new technologies to foster student interaction and collaboration. Dist Educ 27(2):139–153
3. Cheng C-H, Su C-H (2012) A Game-based learning system for improving student's learning effectiveness in system analysis course. Procedia Soc Behav Sci 31:669–675
4. Davenport TH, Jarvenpaa SL, Beers MC (1996) Improving knowledge work processes. Sloan Manag Rev 37:53–66
5. Hummel KA, Hlavacs H (2003) Anytime, anywhere learning behavior using a web-based platform for a university lecture. In: Proceedings of the SSGRR 2003 winter conference, L'Aquila, Italy
6. Kamarainen AM et al (2013) EcoMOBILE: integrating augmented reality and probeware with environmental education field trips. Comput Educ 68:545–556
7. Malone TW (1982) Heuristics for designing enjoyable user interfaces: lessons from computer games. In: Proceedings of the 1982 conference on human factors in computing systems
8. Martono KT, Nurhayati OD (2014) Implementation of android based mobile learning application as a flexible learning media. Int J Comput Sci Issues (IJCSI) 11(3):168
9. Miller LM et al (2011) Learning and motivational impacts of a multimedia science game. Comput Educ 57(1):1425–1433
10. Riley D (2007) Educational technology and practice: types and timescales of change. J Educ Technol Soc 10(1):85–93
11. Schilling MA et al (2003) Learning by doing something else: variation, relatedness, and the learning curve. Manage Sci 49(1):39–56
12. Sharplin E (2002) Rural retreat or outback hell: expectations of rural and remote teaching. Issues Educ Res 12(1):49

13. Tolman EC (1924) The inheritance of maze-learning ability in rats. J Comp Psychol 4(1):1
14. Voogt J, Knezek G (2008) International handbook of information technology in primary and secondary education, vol. 20. Springer Science & Business Media
15. Winn W (2003) Research methods and types of evidence for research in educational technology. Educ Psychol Rev 15(4):367–373

# Smart Stingless Beehive Monitoring System


**C. Edmund and Munirah Ab. Rahman**

**Abstract** The bee honey industry is a very lucrative industry and in just 2015, Malaysia imported RM58 millions ringgit worth of honey products. Although stingless bee keeping seemingly simple, there are a few problems that could affect the failure of success. Stingless bee industry in Malaysia are facing major challenges especially in the respect of the queen, quality inconsistency, low honey production and high price. This is because most of the stingless beehive farm in Malaysia do not have a proper system to monitor and control the environmental parameter of beehive. Moreover, because of the medicinal value of the stingless beehive, there are also a few cases of stolen beehive in Malaysia. Therefore, to prevent these issues, a project named Smart Stingless Beehive Monitoring System (SSBMS) is being developed to monitor and control the environmental parameter of beehive automatically. Sensors are installed at stingless beehive to monitor and control the environmental parameter of stingless beehive such as temperature, humidity, water level as well as the geographical location of beehive through Internet of Things (IoT) platform. The feedback system such as DC fan motor and diaphragm pump have been implemented to ensure the temperature of stingless beehive and water supply for stingless bee can be maintained below 30 °C and in an adequate level respectively. In conclusion, SSBMS is able to help reducing manpower in monitoring the environmental parameter of stingless beehive, gives feedback to maintain the internal beehive temperature and water supply level automatically or based on beekeepers' desires and tracks the geographical location of the stingless beehive through Global Positioning System (GPS).

**Keywords** Smart stingless beehive monitoring system (SSBMS) · Internet of things (IoT) · Global positioning system (GPS)

C. Edmund · M. Ab. Rahman (✉)
Electronic Engineering Department, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussien Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia
e-mail: munira@uthm.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_44

## 1 Introduction

Stingless bees typically can be found in tropical and subtropical region of the world especially in the tropical dry and humid forest [1]. All stingless bees are from the tribe of Meliponini, which has many types of genera, such as Melipona, Trigona, and Heterotrigona. Honey produced by these bee species possesses distinct aroma and taste, slower crystallization process and a more fluidic texture [2].

There are lots of types for stingless bee species, but only 9 species are suitable for bee keeping in Malaysia. There are only 2 species are mostly used for bee keeping for the sake of commercial purpose which is Heterotrigona Itama and Geniotrigona Thoracica since they are very active in the production of honey and propolis. Because of stingless bee honey's medicinal value, stingless bees keeping has been very popular in the tropical and subtropical region of the world.

The same thing happened to Malaysia as it is also keeping stingless bee has only been very popular only in the last decade. The bee honey industry is a very lucrative industry and in just 2015, Malaysia imported RM58 millions ringgit worth of honey products [3]. Beehives are important to be kept in good condition for the working of the bees. Stingless bees prefer to live in an environment that is safe and suitable for survival of the stingless bees. They are however influenced by physical environmental factors that affect their micro-climate. For example, there is an optimum temperature range and relative humidity that various species prefer [4].

The environmental parameter of stingless beehives in Malaysia are mostly being monitored manually. A proper monitoring system for the stingless bee is needed to ensure a conducive environment for the stingless bee habitat. Moreover, stingless bee industry in Malaysia are facing major challenges especially in the respect of the queen, quality inconsistency, low honey production and high price [5]. There are a few cases of stolen stingless beehive in Malaysia. This is due to the high commercial value of stingless bee honey which make it expensive [6].

Therefore, this main goal project is to design a proper integrated sensor system which is SSBMS. To achieve this goal, here are the objectives. Firstly, it is needed to design a stingless beehive monitoring system that can reduce manpower on detecting water supply level, internal humidity, and internal temperature of stingless beehive. Then, it is necessary to develop a feedback system in controlling the temperature and water level of beehive. Moreover, it is needed to develop a GPS tracker on the beehive to be monitored by beekeepers about the geographical location of the beehive by using IoT platform.

## 2 Methodology

In an era of Industrial Revolution 4.0 which means the explosive amount of changes in automation and machine intelligence driven by the combination of software and hardware. Some countries start to realize a concept of "Smart Agriculture" for better

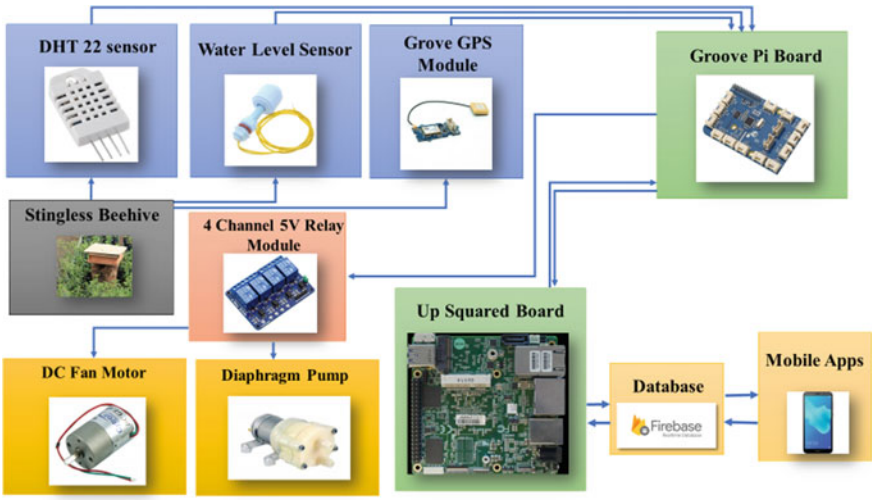



**Fig. 1** System architecture of SSBMS

in efficiency for the control management of agriculture with the growth of telecommunication of 5G network and the application of Internet of Things (IoT) in different field such as apiculture field. Therefore, a more efficient and automatic apiculture system starts to appear left and right in the world recently. Figure 1 shows the system architecture of SSBMS.

Based on the Fig. 1, the sensors which had been used is water level sensor and Grove GPS Module. For Grove GPS Module, it will be used to monitor the geographical location of stingless beehive by beekeepers so that they can keep in track of location of Stingless Beehive even when stingless beehive being stolen or lost. Water level sensor will be used to detect the water supply level of beehives to ensure the water supply. Plus, DHT 22 sensor had been used to detect the internal environment's temperature and humidity of the stingless beehive.

SSBMS operates based on the flowchart in Fig. 2. The data will be collected from the sensors and then send to groove pi board for processing. Then the data will be sent to cloud though Up Squared Board. With the feedback system such as DC fan motor and Diaphragm Pump being implemented, temperature of stingless beehive and water supply for stingless bee can be maintained at below 29 °C and in an adequate level respectively.

With the implementation of Firebase Realtime Database, the data generated by the sensor can be sent 1 data for every 40 s to the platform as a database which is a total of 2160 data per day. It uses the HTTP RESTful API to push and write JSON data generated by the sensors of the system into Firebase Realtime Database. Firebase Realtime Database have no limitation in term of storing the amount of data, so the data can be stored for days, months or even years to be tracked by beekeepers for the understanding of environmental parameter development of stingless beehive. Moreover, the data which stored in Firebase Realtime Database will be collected

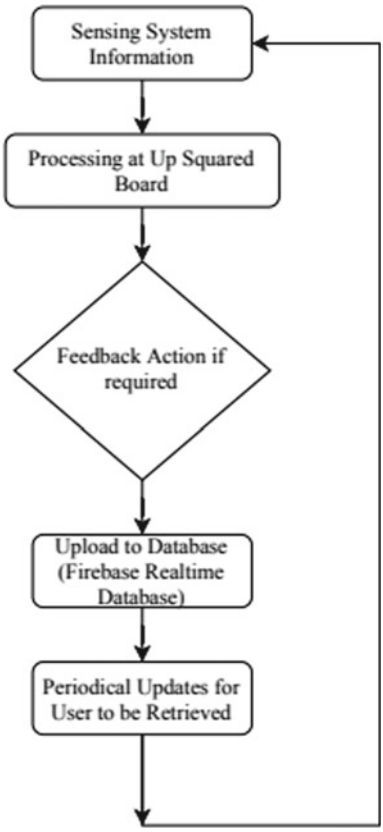



**Fig. 2** Flowchart for operation of SSBMS

by mobile application which created with MIT App Inventor 2 for the purpose of analysis and visualization. Figure 3 shows the schematic diagram for SSBMS.

As for the prototype setup for SSBMS as shown in Fig. 4, the model is separated into three parts which are stingless beehive, water reservoir and box for SSBMS hardware connection. At the side of stingless beehive, DHT22 sensor had been installed to detect the temperature and humidity inside the box through the blue region. The feedback system which is DC fan motor for the mentioned parameters had been installed below the stingless beehive. As for water reservoir, floating sensor had been installed to detect the water level and supply water by diaphragm pump to ensure the water level is always enough for stingless bee. The rest of the hardware connection of SSBMS are being inserted into the box under water reservoir of SSBMS.

As for the software setup, the purpose for software setup is to collect the data from the sensors and send it to the cloud storage which can be accessible to beekeepers and applicable for researchers to make required improvement. All data of SSBMS will be pushed to Google Firebase database in real time. The mobile application gets the required information from the database and illustrates them in graph form. The proposed SSBMS android mobile application has an interactive graphical user

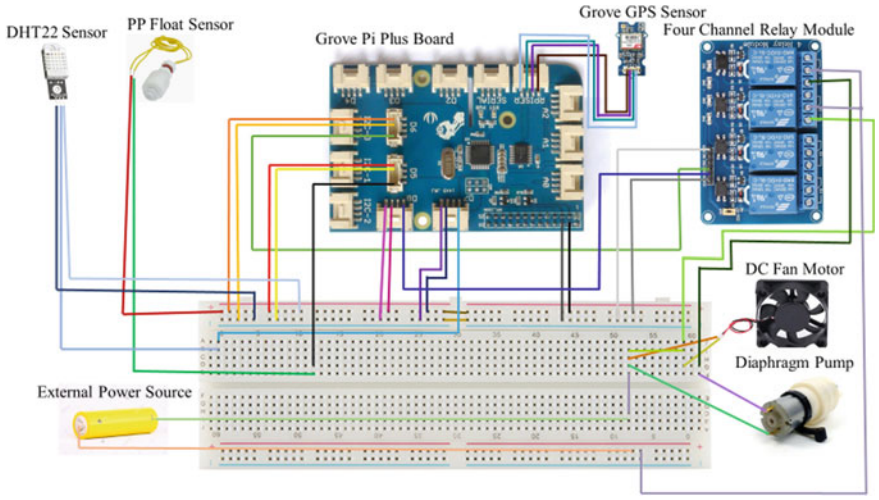



**Fig. 3** Schematic circuit of SSBMS

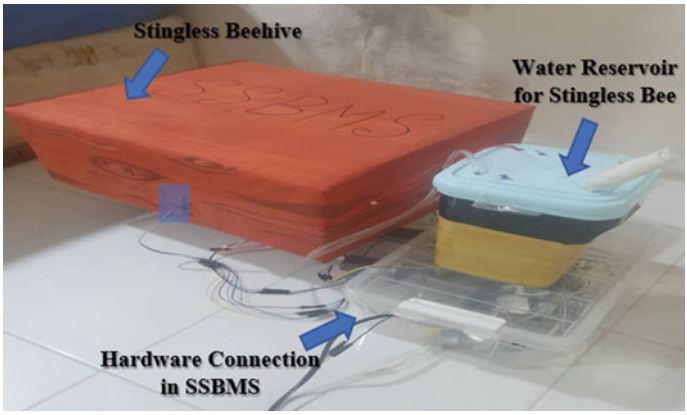



**Fig. 4** Prototype setup of SSBMS

interface (GUI) for user experience to analyze the value of temperature, humidity, water supply level, and GPS coordinate. The data is shown in a graph form and real time. The opening interface of the SSBMS mobile application have a couple of splash pages. The splash pages happen for 5 s to inform user about the overview of the logo and brand before going to the Live Data Chart. The screen name is Live Data Chart as this screen is used to display the sensor data of Smart Stingless Beehive Monitoring System in real time. Moreover, the real time sensor data can be displayed in the form of line chart which can be updated in time of 40 s. The fan motor and the water pump motor can also be monitored and controlled with virtual buttons in the mobile application with a slight delay. Figure 5 shows the page of Live Data Chart.

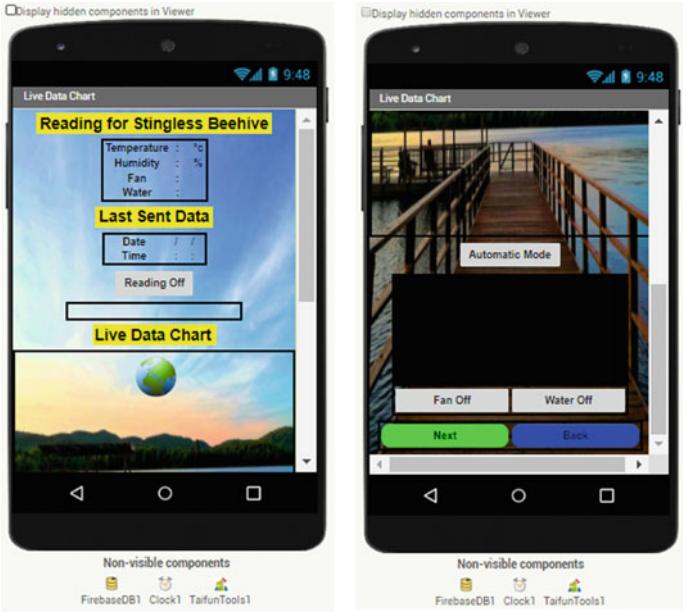



**Fig. 5** Live data chart

After the next button in Live Data Chart had been pushed, it will load into Historical Chart Part 1. This screen is used to display the historical data of Smart Stingless Beehive Monitoring System in the form of line chart. It has a spinner to enable the user to select the mode which are "Minute" or "Hour" to be monitored in. It will keep on update the historical graph data 40 s per time. With the existence of historical graph chart, the user will be able to analyze the development of data which had been extract from Firebase Realtime Database. In "Minute" mode, the time interval for each data is 40 s while for "Hour" mode, the time interval for each data is 16 min. Both of mode has 24 data points to be observed and monitored. Figure 6 shows the Historical Chart Part 1.

As for next screen which is Historical Chart Part 2, this screen is also work the same as Historical Chart Part 1 except for the spinner with modes of "Day" or "Month" instead of "Minute or "Hour". It will also keep on update the historical graph data 40 s per time. In "Day" mode, the time interval for each data is one hour and thirty minutes while for "Month" mode, the time interval for each data is 36 h. "Day" mode has 24 data points while "Month" mode has 30 data points which is meant to be observed and monitored. Figure 7 illustrates the Historical Chart Part 2.

The screen in Fig. 8 also works the same as Historical Chart Part 1 and Historical Chart Part 2 except for the spinner which consists of modes of "1× Year", "2× Years" or "10× Years" instead of "Day or "Month". In "1× Year" and "2× Years" mode, the time interval for each data is 1.5 month. As for "10× Years" mode, it the time interval for each data is 1.5 year. "1× Year" mode has 12 data points; "2× Years"

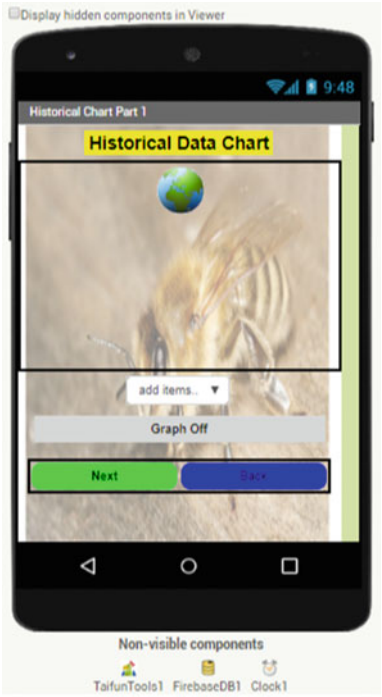



**Fig. 6** Historical Chart Part 1

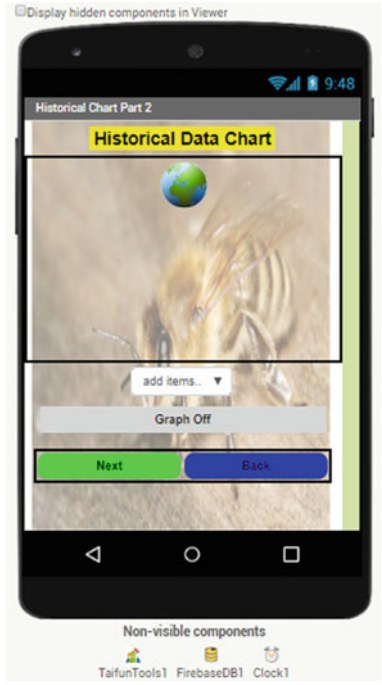



**Fig. 7** Historical Chart Part 2

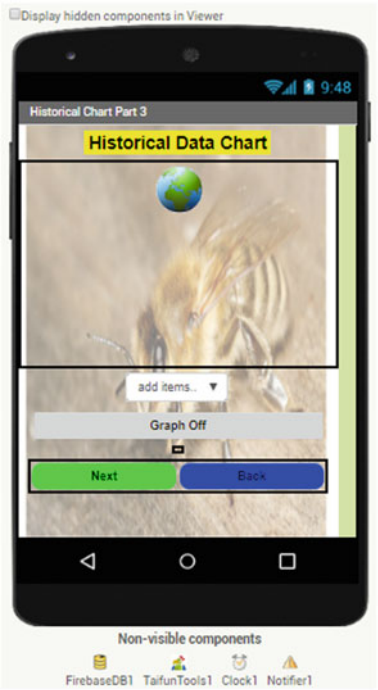



**Fig. 8** Historical Chart Part 3

mode has 24 data points and "10× Years" mode has 10 data points. Figure 8 displays Historical Chart Part 3.

Finally, for the last screen of SSBMS mobile application, its name is SSBMS GPS Tracker This screen is served as a Global Positioning System (GPS) tracker to inform the user which is the beekeeper of Smart Stingless Beehive Monitoring System about the geographical location of beehive. In this SSBMS GPS Tracker, it can be separated into two part. The above part of SSBMS GPS Tracker is the GPS coordinate and address of user's mobile phone whereas the below part of Screen4 will show the GPS coordinate of beehive which installed Smart Stingless Beehive Monitoring System. Screen of SSBMS GPS Tracker had been shown in Fig. 9.

## 3 Results and Discussion

The proposed system was being assembled and tested in a box which are used to imitate the stingless beehive which located in GreenLane Heights, Jelutong, Penang. The box had been implemented with SSBMS and the data were recorded from 18th June 2020 at 7:19 pm to 22nd June 2020 at 12:23 pm. The time for data collection is approximately 89 h and the data are being saved for every 40 s in Firebase Realtime Database. Figure 10 illustrates the actual procedure of pushing the data into Firebase


**Fig. 9** SSBMS GPS tracker

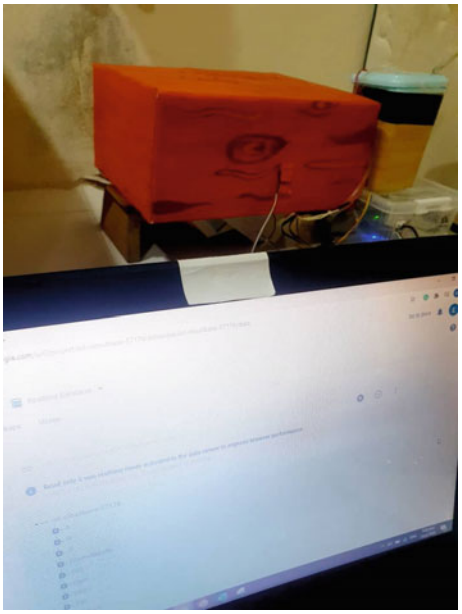


**Fig. 10** The actual procedure of SSBMS pushing data into Firebase Realtime Database at the site

Realtime Database at site.

The number of data which had been extracted from Firebase Realtime Database were a total of 7163 data points to produce output in the line chart form as illustrated in Figs. 11 and 12. Both figures display SSBMS historical data chart regarding the relationship between internal temperature or internal humidity against date and time and the relationship between water or fan against date and time respectively. For the water that is represent as the feedback system for water supply level, it had been recorded as zero from the start of collection of data till the end of data collection. This is because the water reservoir cover is being covered up and it prevent any evaporation of water from it.

Since the time for data collection is just exactly 81 h 26 min and 19 s which starts from 2:42 p.m. of 2nd of July 2020 to 12:08 a.m. of 6th of July 2020, the water supply level had not low enough to be sensed by floating sensor to trigger the diaphragm pump from pumping water. Plus, since the site which had been tested is a controlled environment, no stingless bee had been involved so the water supply level will not be affected.

As for the fan that is used to represent the DC Motor Fan, it had been activated most of the time since the internal temperature for the box that is used to represent Stingless Beehive had been exceed the pre-set threshold which is 30 °C. However, the DC Fan Motor managed to be deactivated at a certain time. The examples of

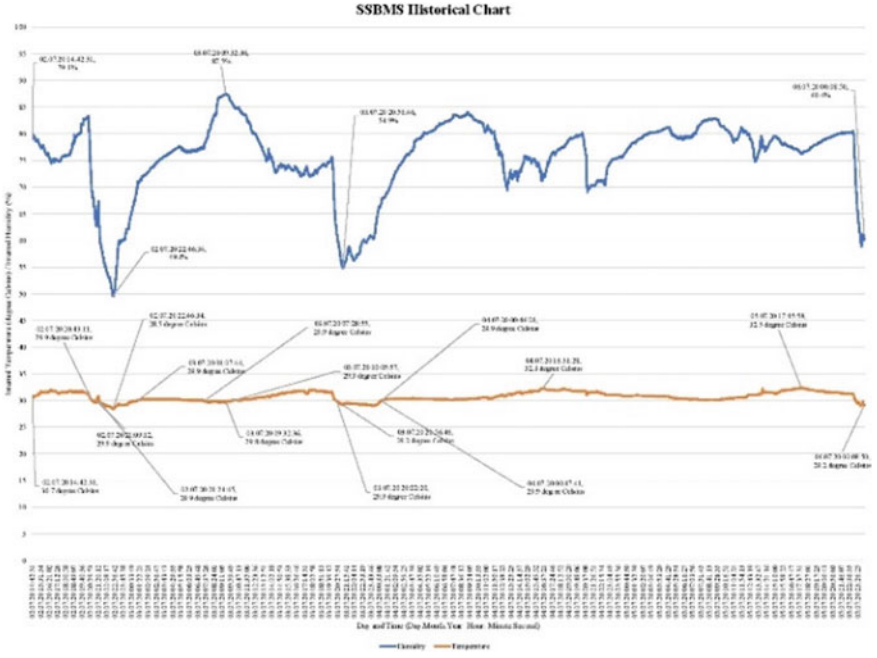



**Fig. 11** SSBMS historical data chart regarding the relationship between water or fan against date and time

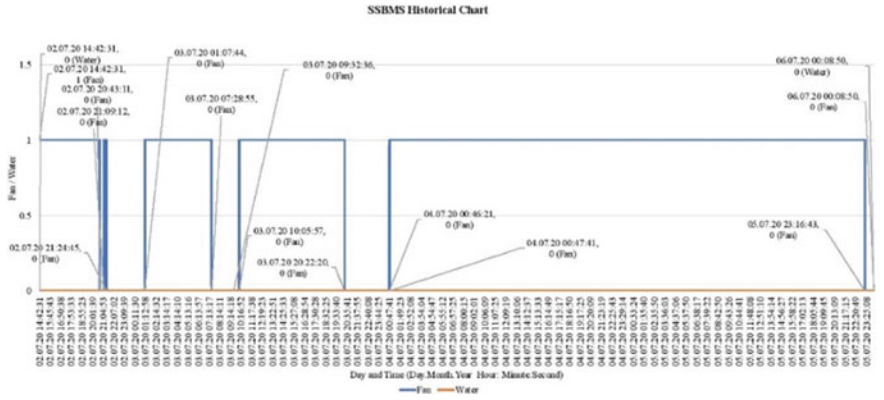



**Fig. 12** SSBMS historical data chart about the relationship between humidity or temperature against date and time

time for fan being deactivated is a few of time periods which are 8:43 p.m. of 2nd of July 2020 to 9:09 p.m. of 2nd of June 2020, 9:24 p.m. of 2nd of June 2020 to 1:07 a.m. of 3rd of June 2020, 7:28 a.m. of 3rd of June 2020 to 10:05 a.m. of 3rd of June 2020, 8:22 p.m. of 3rd of June 2020 to 12:46 a.m. of 4th of June 2020 and finally 11:16 p.m. of 5th of June 2020 to 12:08 a.m. of 6th of June 2020. All of them had a common point which is 29.9 °C since the pre-set threshold temperature to deactivate DC Fan Motor is below than 30.0 °C. The highest internal temperature for the beehive is 32.5 °C at 5.45 p.m. of 5th of July 2020 while the lowest internal beehive temperature is 28.5 °C at 10.46 p.m. of 2nd of July 2020.

When the internal humidity of the box had been recorded during this data collection period, Fig. 11 shows that the internal beehive humidity had been dropped sharply for three times which are 49.6% of humidity, 54.9% of humidity and 60.4% of humidity during 10:46p.m. of 2nd of July 2020, 8:56p.m. of 3rd of July 2020 and 12:08a.m. of 6th of July 2020 respectively. 49.6% of humidity during 10:46p.m. of 2nd of July 2020 is the lowest humidity during this period of data collection. The highest humidity for this data collection period is 87.5% during 9:32a.m. of 3rd of June 2020. When internal humidity of beehive drop, internal temperature will drop as well. Since the site is performed at indoor environment, humid air hold heat more efficiently compared to dry air [7]. Hence, when the internal humidity of beehive drop, internal temperature for beehive will drop too as there is lesser humid in air to hold the heat efficiently.

For the SSBMS GPS tracker, it had been tested during 8:29p.m. of 20th June 2020. After the GPS location and address of user's mobile phone as well as GPS coordinate of stingless beehive had been shown in the mobile application, the yellow button had been pressed and the location can be tracked in Google Map from the current location of user's mobile phone which is Wan Hoi Hotel, Love Lane, Georgetown, Penang at that time and the destination of beehive which located at GreenLane Heights,

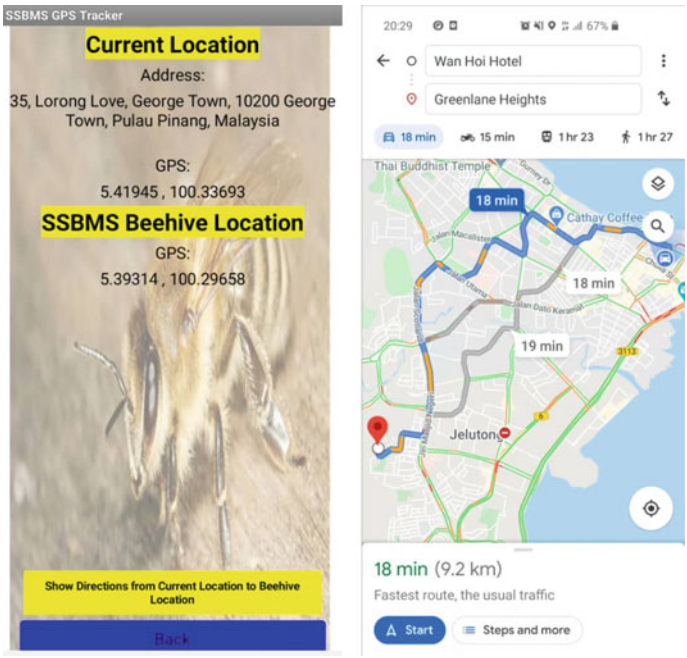



**Fig. 13** SSBMS GPS tracker from Wan Hoi Hotel, Love Lane, Georgetown, Penang to project site, which is GreenLane Heights, Jelutong, Penang

Jelutong, Penang. Figure 13 had illustrated the tracking of project site from location of user.

## 4 Conclusion

As a conclusion, this project had been successfully achieved the three objectives that had been stated. In this paper, it had been presented an integrated sensor system which designed to acquire the selected environmental parameter of stingless beehive which is the internal temperature and internal humidity. Besides that, it can also detect the water supply level of stingless beehive to ensure the water supply for the stingless beehive. Moreover, the system will also provide feedback to maintain the internal beehive temperature and water supply level for the stingless bees. In addition, the system also had GPS tracker installed so that the geographical location of stingless beehive can be tracked at anytime and anywhere.

For future recommendation, it is needed to be tested practically on a real stingless beehive based on yearly period so that the pattern and behaviour of stingless bee can be deeply understood with the help of large scalable yet comprehensive data analysis. Furthermore, the power source for the system should be not completely rely on main

power source. This is because when theft case or blackout happen on the stingless beehive which have the system installed, the system can still be operated by its own without the need of power supply from main power source. Power meter is also needed to be installed on the system to check the power that the alternative power source can be provided to the system without the support of main power source.

**Acknowledgements** The authors would like to thank the Registrar Office and Universiti Tun Hussein Onn Malaysia (UTHM) for providing resources and financial support for this research.

## References

1. Reyes-González A, Camou-Guerrero A, Reyes-Salas O, Argueta A, Casas A (2020) Diversity, local knowledge and use of stingless bees (Apidae: Meliponini) In the Municipality Of Nocupétaro, Michoacan, Mexico. f[Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4061457/. Accessed 18 June 2020
2. Biluca FC, Betta FD, Oliveira GPD, Pereira LM, Gonzaga LV, Costa ACO, Fett R (2015) 5-HMF and carbohydrates content in stingless bee honey by CE before and after thermal treatment. Food Chem 159:244–249
3. Ramli AS, Luqman AH, Basrawi F, Oumer AN, Aziz AA, Mustafa Z (2017) A new cooling technique for stingless bees hive. In: MATEC web of conferences, vol 131, p 03013
4. Ramírez VM, Ayala R, González HD (2018) Crop pollination by stingless bees. Pot-Pollen in Stingless Bee Melittology pp 139–153
5. Ismail MM, Ismail WIW (2018) Development of stingless beekeeping projects in Malaysia. In: E3S web of conferences, vol 52, pp 3–4
6. Malaysia's stingless bee industry faces major challenges. Borneo Post Online, 01-Aug-2019. [Online]. Available: https://www.theborneopost.com/2019/08/01/malaysias-stingless-bee-industry-faces-major-challenges/. Accessed 20 June 2020
7. Air conditioning system: working principle and types with PDF. The Engineers Post, 2020. [Online]. Available: https://www.theengineerspost.com/air-conditioner-working-principle/. Accessed: 21 June 2020

# An Empirical Study to Improve Multiclass Classification Using Hybrid Ensemble Approach for Students' Performance Prediction


**Hasniza Hassan, Nor Bahiah Ahmad, and Roselina Sallehuddin**

**Abstract** Improving machine learning algorithms has been the interest of data scientists and researchers for the past few years. Among the performance problems raised is the classification imbalance issues listed as the top ten. The present study makes comparison and analysis of 5 state-of-art classifiers, 5 ensembles classifiers and 10 resampling techniques for data imbalance. This is done via the used 4413 instances consisting of demographic, economic, and behavioural data from student information systems and e-learning, as well as engineering faculty for the first semester 2017/2018. The use of three sampling types was adapted for the analysis: oversampling, undersampling and hybrid. The experimental results prove to model students' behaviour from e-learning data and bagging decision tree ensemble classifier produces the highest results. Lastly, a hybrid resampling technique, SMOTEENN consistently shows the top result compared to other resampling techniques.

**Keywords** Multiclass classification · Sampling · Students performance prediction · Machine learning

## 1 Introduction

The increase in the amount of data in the education sector has led to the processing of the data to discover knowledge, thus becoming more significant. One of the issues that are frequently given attention by the institutions is to improve student achievement, where the academic information and data are considered to be the most significant predictor in educational data mining.

H. Hassan · N. B. Ahmad (✉) · R. Sallehuddin
Department of Engineering, School of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
e-mail: bahiah@utm.my

H. Hassan
e-mail: nieza1212@gmail.com

R. Sallehuddin
e-mail: roselina@utm.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_45

Conversely, there are several techniques highlighted in studies and frequently used in various domains to overcome these problems, such as resampling, cost-sensitive or ensembles learning. However, there is still lack of studies in literature as the phenomenon is fresh in education domains, particularly to predict students' performance. Despite many previous studies that have been carried out to balance classes in education data, yet more studies have led to binary imbalance problems compared to the multiclass imbalance.

Many studies have been done in the last few decades by researchers, and various prediction models have also been adapted. Besides, some of the previous studies have identified behavioural factors in the use of virtual learning and student demographic factors are among the criteria in the construction of high-performance student performance prediction models [1–4].

In this regard, data preprocessing has been the necessary and crucial steps in data mining and prediction as it ensures the raw data is ready before the next processing. Features selection, noise filtering and data balancing are essential steps in data preprocessing to prepare the information for modelling. However, large amounts of educational data may cause the data to experience noise and other problems, such as class imbalances. Nevertheless, the result is more biased to the majority classes when the data trained using machine learning [5]. As such, may cause the model to tend to the majority class.

Out of the ten problems is the classification imbalance which is the common issue that always occur in data mining [6]. In this regard, researchers have developed various techniques to overcome the imbalanced problem. Nevertheless, most tend to solve binary data imbalance problems than multiclass problems. Therefore, more studies on multiclass imbalance problems need to be implemented to improve performance prediction models [7].

## 2 Related Work

### *2.1 Ensemble Classifiers*

The ensemble classifier is a consolidation of more than one classifier to produce a robust machine learning model for solving classification problems. Several weak learners are united to build a vital model. It is also known as the multiple classification systems, based on several learning models that have used to solve classification problems [8].

A study by [2] proposed a student's performance prediction model based on ensemble techniques with student's behavioural data features. Experimental results show that the consequence achieves improvement when using behavioural features, academic achievement and ensemble methods.

Reference [3] compared different classification methods and develop a students' achievement prediction model by using student's behaviour data and ensemble

methods. By combining Support Vector Machine, Logistic Regression and Random Forest as Majority Vote Ensemble Classifiers, the result shows better performance accuracy.

Another study by [9] compared 7 prediction methods and ensemble technique and increase generalizability as well as the accuracy of the prediction models using feature selection. 3 most successful base methods; Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor was used as ensemble models and had the best results.

Reference [10] compared a combination of data; learning management system, student record system and survey, with state-of-art classifiers; Decision Tree, Support Vector Machine, Artificial Neural Network and ensembles of the 3 base classifiers for students' academic performance prediction. Using multiple data sources along with various ensemble techniques to gain advantageous and high accuracy in student performance prediction.

In a different study by [11], a two-stage model supported by data mining techniques that uses the information available at the end of the first year of students' academic career (path) was proposed to predict their overall academic performance. They discovered that random forests ensembles are excellent to the other classification techniques considered.

Also research by [12] developed a new multimodal perturbation-based ensemble algorithm, RRSB, to improve the performance of ensemble classification. They found that RRSB is robust, with different k values compared with other methods. The experiment shows that it is useful for a big and imbalanced data set.

Reference [13] created a homogeneous ensemble using different classification and regression trees, CART models with modified weighted aging classifier ensemble. The experimental results prove that the result outperformed other machine learning classifiers applied.

## *2.2 Resampling Techniques*

Resampling is the process to overcome the imbalance of class allocation in the training data. The oversampling will duplicate new minority class data [14]. A few examples of oversampling techniques include, Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN) and Random Oversampling (ROS).

On the other hand, the undersampling techniques will eliminate data from the majority class [8]. The few examples of undersampling are Tomek Links (TL), Edited Nearest Neighbors (ENN) and Random Under sampling (RUS). These techniques are more appropriate for the vast size of data. The weakness of undersampling techniques is the data reduction that might cause information loss. However, undersampling and over-sampling techniques have their advantage and disadvantage respectively [15, 16].

Therefore, the combination of these two techniques helps to complement each other. The hybrid technique is the combination of oversampling and undersampling techniques or with ensemble techniques. SMOTE-TL, SMOTE-ENN, SMOTEBoost and RUSBoost are some examples of hybrid approaches. However, the process to create new instances using oversampling techniques may lead to overfitting [5].

### *2.3 Utilization of Resampling Technique for Imbalanced Data*

Data level and algorithm level are 2 class imbalanced classification approach used to manage and solve the imbalance problems. The data level classification approach in sampling helps the ratio of classes to improve and balance. In the algorithm level of classification approach, classifiers utilized to enhance the learning task by fine-tuning [14].

Interesting work by [5], discovered that 4 classifiers show better performance when combining the proposed two noise filter method with class imbalance approach. They developed students' performance prediction using two datasets effected by noise and class imbalanced data. In the first experiment, the behaviour of five class imbalance techniques analysis to oversample the datasets; SMOTE, SMOTE-ENN, SMOTE-TL, ROS and weighting using 4 classifiers; Support Vector Machine, Logistic Regression, Random Forest and AdaBoost was observed. The second experiment proposed two new techniques, BST-CF, BST-EF on different classifiers to identify a set of noise that should remove.

Reference [16] performed an empirical study of class noise and imbalance. 11 different classifiers analyzed on sample data to predict the quality of the software. By using 12 real-world datasets from software quality data and seven sample techniques, they discovered and proposed a new SMOTE-IPE technique, that proveed to tackle noise and borderline of instances in imbalanced data sets problem.

A study by [17] proposed Diversified Error-Correcting Output Codes (DECOC), that is a novel multiclass imbalanced classification. This method has proven to give a significant improvement in accuracy compared to 17 other classifiers.

Research by [18] focuses on improving the true positive rate of minority class (GDM). About 886 instances was collected, which contain patients with diabetes mellitus disease information. In which two low undersampling techniques (RUS and NCL) was also added to balance the dataset before classification. This process involves undersampling the majority class, which balances the dataset before the classification was applied. The 3 learning algorithms that was adapted are; Decision tree (pruned), Decision tree (unpruned) and RIPPER.

Another study by [19] examined multiclass imbalanced problems, such as classes overlapping, shortage of data, and mixed various types of data. This approach created covariance in minority class and create synthetic samples. Experiments was executed using two datasets and 15 multiclass imbalanced benchmarks. Additionally, their

study introduced GSVD (Generalized Singular Value Decomposition) in Adaptive Mahalanobis Distance-based Over-sampling (AMDO) and manage to resolve problems with mixed-type of data, develop a part of balanced resampling scheme and utilize sample synthesis.

In another study by [20], an error diagnosis model was improve by proposing the knowledge-based procedures. This was carried out via 2 different feature selection processes to decide on the essential set of features. Some ensembles utilized in training the model. Their observations show the importance of performing the class imbalance techniques as it can reduce the mispronunciation of incorrect products.

Research by [21] finds out that Kernelized extreme learning machine (KELM) produces better scores compared to the traditional extreme learning machine (ELM), that uses random input parameters. The researchers proposed a generalized CSKELM (GCSKELM), the extension CSKELM, that enable to solve the multiclass imbalanced problems directly.

Reference [11] developed dynamic ensemble selection for multiclass imbalanced datasets (DES-MI) where the candidate classifiers were assessed with weighted instances in the neighborhoods. While also 2 components were proposed in the DES-MI T to balanced training datasets and selection of suitable classifiers. Thereafter, a preprocessing procedure was developed to balance the training data and the weighting mechanism was utilized to highlight the most powerful classifiers.

## 3 Methodology

### *3.1 Datasets*

The present study is a combination of two data sources; the student information system (SIS) and e-learning (EL). Students' behaviour of using e-learning is processed to integrate with the student information system. The data employed was gathered from a Malaysia public research university consist of 4413 students from the Faculty of Engineering in the first session of 2017/2018 were used [22].

### *3.2 Tools*

Python and Jupyter Notebook are programming and software used to run the experiments. This was runs in the Windows platform with Intel® Core™ i7 and 16 GB of RAM. Furthermore, the present study used the Scikitlearn python package for data preprocessing and machine learning algorithms while the Imblearn python package was used for imbalance class resampling.

### 3.3 Preprocessing

Pre-processing is a data mining step to enhance the data quality and ensure the modeling of machine learning to becomes more efficient. It is also essential to provide the best data generation for modelling using machine learning methods. Scikitlearn is the preprocessing package in python that helps to perform a sequence of preprocessing. Firstly, an initial feature filtering is carried out where unwanted data are eliminated. Afterwards, a data transformation execution in which the categorical data is transformed into numerical values to prepare all for modelling is carried out. Finally, the data normalization is performed to change all the data figures to a small range. This is important to make data ready for the modelling process.

More so, the selection of the important features is very crucial to improve predictive performance. This is to eliminate irrelevant, redundant data and ensure the quality and accuracy of the learning model developed. The presence of unnecessary features can decrease training speed and generalization performance of test data. Feature selection reduces data dimensions to increase data mining results. For the present study 19 features were selected in the first stage of filtering. They were classified into 3 features types: academic background, demographical with socioeconomic and behavior e-learning.

Through this study, 10 features were selected using 6 feature selection algorithms: 3 wrappers and 3 filter techniques with statistical evidence. The steps include, sorting together with a combination of the wrapper and filter algorithms. The results show 7 behavioural e-learning features, 1 economic feature, 1 academic feature and 1 demographic feature among the top 10 ranking features that are important to use as predictors in the student predictive model. The features (bold) are as shown in Table 1.

According to [2], students' behaviour features contribute to the improvement of accuracy in the model proposed. They discovered a strong relationship between academic and behaviour features. Behaviour features improved the performance accuracy in students' performance prediction models [10, 23]. A finding by [24] proves a strong relationship between students' behaviour and academic performance.

### 3.4 Performance Measure

There are 3 classification performance metrics used in the empirical works of the present study:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

**Table 1** Features and description [33]

| Category | Feature | Description |
|---|---|---|
| Academic background features | Study method | Coursework or research |
| | **Programme** | **Programme of study** |
| | CGPA | Cumulative grade point average |
| | Year_Intake | Year of students' intake |
| | Education_Mode | Part time/full time |
| Demographics/socioeconomic features | **Family_Income** | **Family income range** |
| | Student_Status | Student status |
| | **Scholarship** | **Name of scholarship** |
| | **Gender** | **Student gender** |
| | **Age** | **Student age** |
| | **Nationality** | **Student nationality** |
| | Disability | Disability status |
| Behaviour features (e-learning) | **User_Loggedin** | **Count no of login** |
| | **Course_Viewed** | **Count course viewed** |
| | **Course_Module_Viewed** | **Count resources viewed** |
| | **Discussion_Viewed** | **Count forum/discussion viewed** |
| | **Submission_Form_Submitted** | **Count course submitted** |
| | **Attempt_Viewed** | **Count assignment viewed** |
| | **Assesable_Submitted** | **Count assignment submitted** |

$$FScore = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

## *3.5 Modelling*

For training the machine learning model, the data were split into two parts, train data and test data with ratio 70: for training and 30 for testing. Thereafter, the cross-validation technique with 10 folds was applied to every classification model. Also, grid search was utilized to get the best hyperparameters for every model.

# 4 Empirical Experiment Result and Analysis

## *4.1 Experiment I*

Table 2 shows the results of the empirical study using the F-score performance metric after applying the 10 resampling techniques. The experiment trained models using 5 base classifiers and 5 ensemble classifiers for the top ten best features. The base classifiers are; Decision Tree (DT), K-Nearest Neighbour (KNN), Neural Network (NN), Naïve Bayes (NB) and Support Vector Machine (SVM). AdaBoost (AB), Bagging (BGG), Random Forest (RF), Gradient Boosting (GB) and XGBoost (XGB) are ensembles classifiers used to evaluate the performance.

Among the classifier trained, all show improvement after balancing the multi-class labels. Bagging shows the highest accuracy result among all classifiers. The hybrid technique, SMOTEENN is the most suitable imbalance resampling technique used for the education data since it shows the most consistent highest result with all the classifiers employed. This indicates that SMOTEENN is the most suitable hybrid technique in the present study.

From the first experiment executed, Bagging achieves the highest F-score while sampling using SMOTEENN consistently shows the highest result among all 10 resampling techniques used.

**Table 2** Applied imbalance technique

| Type | Sampling | DT | NB | NN | SVM | KNN | RF | BGG | AB | GB | XGB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | No sampling | 0.411 | 0.389 | 0.358 | 0.278 | 0.412 | 0.411 | 0.417 | 0.415 | 0.469 | 0.425 |
| Over sampling | SMOTE | 0.664 | 0.449 | 0.57 | 0.478 | 0.677 | 0.726 | 0.73 | 0.661 | 0.731 | 0.716 |
| | ROS | 0.779 | 0.455 | 0.552 | 0.485 | 0.64 | 0.779 | 0.775 | 0.783 | 0.69 | 0.674 |
| | ADS | 0.663 | 0.436 | 0.555 | 0.452 | 0.65 | 0.708 | 0.706 | 0.669 | 0.705 | 0.702 |
| Under sampling | RUS | 0.453 | 0.44 | 0.481 | 0.376 | 0.5 | 0.406 | 0.471 | 0.476 | 0.586 | 0.595 |
| | NM | 0.678 | 0.438 | 0.565 | 0.47 | 0.664 | 0.733 | 0.725 | 0.751 | 0.738 | 0.716 |
| | CNN | 0.786 | 0.608 | 0.727 | 0.575 | 0.822 | 0.822 | 0.828 | 0.778 | 0.8 | 0.783 |
| | ENN | 0.552 | 0.562 | 0.559 | 0.294 | 0.568 | 0.646 | 0.586 | 0.547 | 0.647 | 0.675 |
| | TL | 0.443 | 0.44 | 0.406 | 0.282 | 0.44 | 0.474 | 0.466 | 0.432 | 0.465 | 0.434 |
| Hybrid | SMOTEENN | 0.789 | 0.609 | 0.721 | 0.581 | 0.823 | 0.831 | 0.855 | 0.792 | 0.801 | 0.787 |
| | SMOTETL | 0.659 | 0.452 | 0.581 | 0.481 | 0.67 | 0.732 | 0.724 | 0.667 | 0.725 | 0.702 |

**Table 3** Bagging various base classifiers

| Sampling | BGG-DT | BGG-NB | BGG-KNN | BGG-SVM |
|---|---|---|---|---|
| SMOTEENN | 0.855 | 0.629 | 0.841 | 0.706 |

### *4.2 Experiment II*

The second experiment utilized the result from the first experiment where the Bagging gain the highest F-score and SMOTEENN and shows consistency among all resampling techniques used. Table 3 shows the experiment of Bagging ensemble learning using multiple base classifiers and balance of the data using SMOTEENN.

The experiment above shows that among the base classifiers utilized, Bagging with decision trees gain the highest F-score result, followed by Bagging with k-nearest neighbours.

Reference [25] proposed an ensemble meta-based tree model (EMT), to predict student performance. EMT is a combination of 2 consistent machine learning classifiers. The results show the NBTree and Adaboost_J48 classifiers outperform and has applied in the EMT. It shows the based classifier with tree types perform well to implement to ensemble classifiers. Also, a study by [24] proposes the bootstrap aggregating technique on the hybridization of two base classifiers; Naıve Bayes (NB) and Decision Tree (DT). Conversely, NB was utilized to remove noise instances from the training set, then follow by train the DT. The hybrid Bagged Naıve Bayes-Decision Tree (BNBDT) hybrid algorithm shows the best result to improve the classification accuracy of various multiclass problems. In another study, [26] applied and integrated SVM base classifier with Bagging classifier and proposed a novel method for druggable proteins prediction. The experimental result shows that feature selection, feature extraction, ensemble learning, and other algorithms improved the model significantly.

While for education domain, research by [2, 3, 27–32] proved that ensemble learning classifiers produce students' performance prediction model improvement and better accuracy.

## 5 Conclusion and Future Work

Student performance is an important factor needed to be prioritized as it indirectly reflects the achievements of educational institutions. In the education sector, academic achievement is a fundamental factor that needs to be prioritize and monitor consistently from time to time. Student data consisting of demographics, socioeconomics, students' profile and attitude of online learning contained valuable information that can be interpreted as knowledge. However, the educational sector needs more research to discover additional knowledge and build accurate student performance prediction models using the hybrid of ensembles.

Future research is recommended to be focus more on the improvement of the bagging technique where hybridization of classifiers improves the multiclass issues. Also, cross-domain data samples need to be used to train the hybridization of bagging. Furthermore, future work needs to concentrate more on the hybrid of bagging techniques with various base classifiers and hyperparameters tuning, since bagging has proven to produce the highest result.

**Acknowledgements** The authors are grateful to the Ministry of Education and Universiti Teknologi Malaysia for the supplied of data used in the present study.

## References

1. Adejo O, Connolly T (2017) An integrated system framework for predicting students' academic performance in higher educational institutions. Int J Comput Sci Inf Technol 9(3):149–157
2. Amrieh EA, Hamtini T, Aljarah I (2016) Mining educational data to predict student's academic performance using ensemble methods. Int J Database Theory Appl 9(8):119–136
3. Salini A, Jeyapriya U, College SM, College SM (2018) A majority vote based ensemble classifier for predicting students academic performance. Int J Pure Appl Math 118(24):1–11
4. Cerezo R, Sánchez-Santillán M, Paule-Ruiz MP, Núñez JC (2016) Students' LMS interaction patterns and their relationship with achievement: a case study in higher education. Comput Educ
5. Radwan AM, Cataltepe Z (2017) Improving performance prediction on education data with noise and class imbalance. Intell Autom Soft Comput 8587:1–8
6. Qiang Y, Xindong W (2006) 10 Challenging problems in data mining research. Int J Inf Technol Decis Mak 5(4):597–604
7. Wang S, Yao X (2012) Multiclass imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cybern Part B Cybern
8. Gudivada VN, Irfan MT, Fathi E, Rao DL (2016) Cognitive analytics: going beyond big data analytics and machine learning. Handbook of Statistics
9. Marbouti F, Diefes-Dux HA, Madhavan K (2016) Models for early prediction of at-risk students in a course using standards-based grading. Comput Educ 103:1–15
10. Adejo OW, Connolly T (2018) Predicting student academic performance using multi-model heterogeneous ensemble approach. J Appl Res High Educ 10(1):61–75
11. Miguéis VL, Freitas A, Garcia PJV, Silva A (2018) Early segmentation of students according to their academic performance: a predictive modelling approach. Decis Support Syst 115:36–51
12. Zhang Y, Cao G, Wang B, Li X (2019) A novel ensemble method for k-nearest neighbor. Pattern Recognit 85:13–25
13. Mienye ID, Sun Y, Wang Z (2020) An improved ensemble learning approach for the prediction of heart disease risk. Inf Med Unlocked 20
14. Blagus R, Lusa L (2013) SMOTE for high-dimensional class-imbalanced data. BMC Bioinf
15. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2010) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man, Cybern Part A Syst Hum 40(1):185–197
16. Seiffert C, Khoshgoftaar TM, Van Hulse J, Folleco A (2014) An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Inf Sci (Ny)
17. Bi J, Zhang C (2018) An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. Knowl-Based Syst 158:81–93

18. Folorunso SO, Adeyemo AB (2013) Alleviating classification problem of imbalanced dataset. Afr J Comput ICT 6(2):137–144
19. Gopalakrishnan A, Kased R, Yang H, Love MB, Graterol C, Shada A (2018) A multifaceted data mining approach to understanding what factors lead college students to persist and graduate. In: Proceedings of computing conference (2017)
20. Han S, Choi HJ, Choi SK, Oh JS (2019) Fault diagnosis of planetary gear carrier packs: a class imbalance and multiclass classification problem. Int J Precis Eng Manuf 20(2):167–179
21. Raghuwanshi BS, Shukla S (2019) Generalized class-specific kernelized extreme learning machine for multiclass imbalanced learning. Expert Syst Appl
22. Katuwal R, Suganthan PN, Zhang L (2018) An ensemble of decision trees with random vector functional link networks for multi-class classification. Appl Soft Comput J 70:1146–1153
23. Francis BK, Babu SS (2019) Predicting academic performance of students using a hybrid data mining approach. J Med Syst 43(6)
24. Singh Namrata SP (2019) A novel Bagged Naıve Bayes-decision tree approach for multi-class classification problems. J Intell Fuzzy Syst, p 2261
25. Almasri A, Celebi E, Alkhawaldeh RS (2019) EMT: ensemble meta-based tree model for predicting student performance. Sci Program
26. Lin J, Chen H, Li S, Liu Y, Li X, Yu B (2019) Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier. Artif Intell Med 98:35–47
27. Yang X, Kuang Q, Zhang W, Zhang G (2017) AMDO: an over-sampling technique for multi-class imbalanced problems. IEEE Trans Knowl Data Eng 30(9):1672–1685
28. Nam SJ, Frishkoff G, Collins-Thompson K (2017) Predicting students' disengaged behaviors in an online meaning-generation task. IEEE Trans Learn Technol 1382:1–14
29. Zollanvari A, Kizilirmak RC, Kho YH, Hernandez-Torrano D (2017) Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. IEEE Access 5:23792–23802
30. Athani SS, Kodli SA, Banavasi MN, Hiremath PGS (2018) Student performance predictor using multiclass support vector classification algorithm. In: Proceedings of IEEE international conference on signal processing and communication, ICSPC (2017), vol 2018, pp 341–346
31. Sun Z, Sun L, Strang K (2018) Big Data analytics services for enhancing business intelligence. J Comput Inf Syst 58(2):162–169
32. Iam-On N, Boongoen T (2017) Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. Int J Mach Learn Cybern 8(2):497–510
33. Hassan H, Ahmad NB, Anuar S (2020) Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. J Phys: Conf Ser

# A Review on Deep Learning Approaches to Forecasting the Changes of Sea Level


**Nosius Luaran, Rayner Alfred, Joe Henry Obit, and Chin Kim On**

**Abstract** The amalgamation of atmospheric elements indicates positive trends in sea level rise which has had a significant impact on nearly 60% of the world's population living in the low elevated coastal area. In this paper, we first discuss potential factors leading to the rise in sea level and negative impacts on future development along the coastal region. Then, methods of acquiring sea level data which revolutionize the study of variation at sea level will also be reviewed and discussed. The present paper aims to review several Deep Learning (DL) algorithms that address critical issues of forecasting, specifically a time variable known as time series by managing complex patterns and inefficiently capturing long-term multivariate data dependency. Asynchronous data handling required correct theoretical framework processes. Based on the review conducted, the deep learning architecture is capable of generating accurate prediction at sea level which can be used as decision-making tools for managing low-lying coastal areas.

**Keywords** Deep learning · CNN · LSTM · GRU · Non-astronomical · Sea level trend

## 1 Introduction

The prediction of Long-term changes in sea level becomes more challenging because of the unpredictable global trend of non-atmospheric components. Once it comes to

N. Luaran (✉) · R. Alfred · J. H. Obit · C. K. On
Knowledge Technology Research Unit, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
e-mail: nosiusluaran@gmail.com

R. Alfred
e-mail: ralfred@ums.edu.my

J. H. Obit
e-mail: joehenry@ums.edu.my

C. K. On
e-mail: kimonchin@ums.edu.my



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_46

changes in the level of the sea, most experts believe that the effect of atmospheric elements has affected the frequency and variability of tides and sea levels across the world from coast to coast. Nevertheless, local non-astronomical sources such as temperature, salinity, current, seawater pressure, wind, seawater topography and local water depth [1, 2] are also factors and parameters that affect the accuracy of the tidal analysis and prediction. National Oceanography Center, 2016 emphasizes the significance of considering metrological effects due to static atmospheric pressure data and dynamic wind stress induced by storm surges (positive or negative surges) to the normal tidal level on the open coast.

Global Mean Sea Level (GMSL) variation shows positive trends [3] and which can lead to significant increase [4]. A number of researchers [5–7] have recently indicated an extreme sea level in several places around the world over the last few decades which are approximately in line with the International Penal on Climate Change (IPCC) 5th Assessment Report (AR5) model projection. The Satellite Altimetry observations gathered from TOPEX/Poseidon, Jason-1, Jason-2 and Jason-3 estimated GMSL the sea level rise due to climate change for the last 25 years is $0.084 \pm 0.025$ mm/y$^2$ [8]. The global measurements related to the global average sea level have risen by approximately 8–9 in. (21–24 cm) since 1880, with approximately 8 cm (3 in.) occurring since 1993 and recent work indicates that the largest possible contributors to rising sea levels are Greenland and Antarctic [9–13]. In addition, [14] also highlights the importance of spatial variability, especially in ocean density, ocean mass redistribution, ocean mass change, related gravitational effects and vertical land motion. Furthermore, it indicates that rapid sea level changed due to rapid changes in ice sheet dynamics [8].

The rising sea levels have significant impacts to approximately where 60% of the world's population living in the low-elevation coastal zone [4] and islands region [15] and expected to continue rising for centuries. The Sea level rise adversely impact the sustainability development of coastal areas as they directly impact the physical property, important coastal infrastructure including transportation and utility infrastructure, production process, natural and cultural resources, and indirectly, impact the ecosystem, economic activities, income, wealth, public health, safety, and overall social well-being of the coastal community [4, 16].

## 2 The Fundamental of Tidal Shift

The fundamental theory of tidal shift in the sea level can be demonstrated as a regular fluctuation in the sea level which primarily influence by the astronomical components such as gravitational forces exerted by the moon, sun and the rotation of the earth. Tidal periodic phenomena were derived and discovered on the basis of Newton gravitational theory which can be denoted in the form of sinusoidal function which basically has three parameters, namely frequency, amplitude and phase [17].

Multiple researchers agreed that the acceleration of GMSL based on the basis of reconstructed model using Satellite Altimetry data indicates positive inclination [8,

18]. However, [9] noted that the Global Mean Sea Level (GMSL) acceleration from the earlier twentieth century was un-predictable which inferred from Tide Gauge (TG) records.

## 2.1 The Impacts of Sea Level Rises in Malaysia

In Malaysia, the increase in Sea level is one of the most alarming and costly consequences of climate change affecting sustainable growth in coastal regions which impacting directly on substantial coastal infrastructure, physical properties and indirectly, the coastal community's environment, economic activities, employment, public health, protection and overall social well-being [16]. A moderate change in mean sea Level would lead to a significant increase in the number of extreme water levels [19].

The National Hydraulic Research Institute of Malaysia (NAHRIM) recorded an average increase of 2.73–7.0 mm (per year) in Sea Level Rise (SLR) over 30 stations in Malaysia from 1993 to 2010 and found that there has been a substantial increase in SLR trends over the last 5 years compared to the SLR trends over 20 years ago [20]. Consequently, if the ice shelves in Greenland and Antarctica begin to increase heat and marine melting, it is difficult to prevent disintegration of large-scale ice sheets that will cause sea level to rise.

It is therefore important to introduce a state-of-art sea-level forecasting techniques that have the capacity to provide realistic forecasting under non-extreme conditions (i.e., storm surges are not captured) and also to support routine coastal decision-making and provide a benchmark for potential Mean Sea Level forecasts. However, the tide height shift are depends directly on the two components that is long-term astronomical and short-term meteorological impact due to extreme weather conditions [21]. In addition, several sources of datasets derived from tide gauges, satellite altimetry, Global Navigation Satellite System (GNSS), leveling campaigns, meteorological stations can be used to investigate the sea level variation [22].

## 2.2 Tidal Data Observation

This section will briefly describe the key sources for tidal heights datasets. In the last decades, Satellite Altimetry era, NASA's remote sensing satellite altimetry measurement has revolutionized our sea-level awareness, ocean circulation, and climate variability studies [23]. Satellite Altimetry information has been utilized widely to investigate the Spatial-Temporal behavior [24], Amplitude's Nodal Modulation [3], potential effects of ocean level variability and also for geoid determination and current circulations study [16]. Satellite Altimetry (SA) specifically calculates sea-surface elevation in the gridded block by interpolation and offers high-precision of about 1 mm/year global sea-level change [24, 25]. Satellite altimetry mission, TOPEX,

JASON1, JASON2, ERS1, ESR2 and ENVISAT appears very promising particularly for studying sea level, as it provides broad coverage for measuring sea level [23].

However, the accuracy of direct observational techniques using Satellite Altimetry to measure the Sea Surface Height (SSH) decreases dramatically along coastal area due to near coast effect [24]. The near coast effects caused the accuracy significantly lower along the coast compared in deep sea. Therefore, Tide Gauge (TG) data played more important role in determining the characteristics and behavior along the coastal area and gives higher precision of sea level change along coast [23–26]. TG data recorded using conventional tide gauges, automatic tide gauges and water level recorders.

### *2.3 Forecasting the Rise of Sea Level*

The prediction of rising sea levels is becoming more critical due to erratic trends, development and impacts such as the prominent effect of Global warming which includes the ice melting and glacier significantly accelerate sea level rise global. The forecasting of sea levels at different geographical locations is necessary for any analysis to recognize and mitigate serious consequences. As several studies have suggested, changes in the level of the sea impact the marine environment, submergence of islands, changes in microclimates, coastal ecosystems, movement of species. A range of industries that rely on sea level information for harbor planning and decision-making, near shore marine firms, hydrographic surveyors, oil exploration/exploitation activities and so on. Therefore, advance planning for future development especially along coastal area require sea level evidence to identify dynamical models of sea level patterns [27].

## 3 Deep Learning Approach to Tidal Level Forecasting

The irregular rising of sea level is become the major challenge for society in the twenty-first century as reported by International Penal on Climate Change (IPCC) [23]. Computational intelligence approaches become actively used to as forecasting method for sequential dataset. High precision of future sea level forecasting need reliable dataset and valuable information that affecting the sea level variations. Non-Astronomical components that are the meteorological mechanisms are the most important factor in establishing model for tidal prediction. Principally, Deep learning modules employ number of layers to solve problem on related time series from the same domain due to the increase volume of data and related information which is considered valuable information affecting the sea level variations.

This section presents the overview of Traditional Univariate Tidal Forecasting methods, Machine Learning Method and Deep Learning approach in forecasting future sea level height change.

## *3.1 Traditional Univariate Tidal Forecasting Method*

Traditional Univariate methods used for tidal forecasting tools are Exponential Smoothing State Space Model (ESMs) and Auto Regressive Integrated Moving Average (ARIMA) as quantitative forecasting methods for sea level rise, using model level, trend and seasonal decomposition [28]. Traditional univariate techniques that is consider individual time series are suitable to use with minimal volume of dataset and unable to exploit multi-variate information [29]. This mean that, less number of parameters are used in this techniques compared to other complex machine learning methods. This method is incapable of extracting valuable knowledge and to deliver high quality prediction. In General, this method is extensively employed as an benchmark result to other Neural Networks and computational intelligence methods [30].

In general, the application of Traditional method for forecasting tidal height, such as Harmonic Analysis (HA), which incorporated tidal characteristics of the many sinusoidal constituents and employed least squares adjustment can achieved adequate result to support marine activities [27]. The enhancement approach, using the combination of conventional method, has also enhanced the overall prediction and increased the accuracy of the harmonic analysis [31]. Nevertheless, the outcome of the prediction is based only on changes in single tidal amplitudes due to its limitation to inter-correlation with other external factors that directly influenced the changes in tidal height such as non-astronomical components and others. This is becoming the main challenges of using the traditional method to predict future tidal height patterns.

## *3.2 Artificial Neural Network (ANN)*

Artificial Neural Networks (ANNs) method for time series prediction since 1990s has become viable, leading to the publication of considerable amount of literature. The basic ANNs model has been used extensively since 1990s for approximation of non-linear time series, with acceptable prediction of water level prediction. ANNs used for sea level prediction and assessing the performance of other techniques [32, 33].

Artificial Neural Network (ANNs) established using a mathematical model that can be trained to tackle problems through machine-learning neurons and also its capability of extracting direct knowledge from complex non-linear relationship between the inputs and outputs which is one of the main contributions in machine-learning approach [34]. The inputs and outputs interconnection illustrated in this mathematical form:

$$Y = f(X_n) \tag{1}$$

where, $Y$ is the output vector and $X_n$ is the n-input array containing of variables $X_1$, …, $X_n$. The number of neurons in each Hidden layers, bias factor and activation function of the ANNs are the components of Black-Box that execute the function to solve specific problem. However, ANN conventional model has significant limitation in solving sequential data such as tidal variation over time factor.

Some of the challenges in the application of machine learning tools namely ANNs to achieve an adequate sea level prediction are to capture non-linear and complex underlying characteristics of physical phase to forecast and define better solution of tidal problems. Therefore, some advanced mathematical techniques are required to resolve the lack of aforementioned observation. Machine learning model utilize, assimilate and learn from evidence of past climate trends using observational dataset to predict the future development [35]. Nevertheless, Machine Learning becomes more useful when insufficient data for executing process-based modeling and without performing downscaling operation.

ANNs enable the tidal analysis composed by several components factor, which are the disturbance factors of tidal level, to ameliorate the accuracy of tidal investigation and improve the low accuracy of a single harmonic analysis [36]. ANNs can be viewed as robust computational method capable of exploiting the non-linear and complex underlying characteristics of physical process with level degree of accuracy. Furthermore, diverse mathematical techniques and amalgamation of various models were established to search for an adequate sea level forecasting methods. However, ANN conventional model has significant limitation in solving sequential data such as tidal variation over time factor and its time learning requirement.

### *3.3 Deep Learning Approach in Tidal Prediction*

In this section, we briefly describe various Deep Learning methods that extensively used algorithms for future forecasting of the time series datasets in different fields. Deep learning methods have had an exceptional effect on the solutions to a wide variety of problems [37]. There are three most popular and efficient Deep Learning networks used to tackle critical forecasting problems contains a time component known as time series. Explicitly, there are Recurrent Neural Network (RNN) and enhanced RNN that is Long Shot Time Memory (LSTM) and Gated Recurrent Unit (GRU) which work-well really well [38].

RNN architecture provide extensive empirical analysis of forecasting and conclude that RNNs is works well in seasonality modelling directly by skipping the homogeneous seasonal patterns-based deseasonalization process [29]. Further empirical evaluations have indicated that RNNs has strong capability in three task groups: arithmetic computation, XML modeling and language modeling.

RNNs are extension of Neural Networks for sequential datasets [38]. In general, RNNs mechanics used the input data to derive the current prediction along with the

outputs. The repeating modules functions of RNN which stored important record from previous steps is successfully used in learning sequences. Nonetheless, there are many difficulties in applying RNNs time series data for future forecasting as the model is not easy to train and cannot be retrieved from previous results. This implies that RNNs do not have the capacity to store forward moving values in time. The main limitation of RNNs is the Vanishing Gradient Problems which happened when the gradient become too insignificant or huge.

Long Short-Term Memory (LSTM) networks are an ideal model used for sequential tidal data due to its capability to learn long-term dependencies time-series data [39]. Furthermore, LSTMs is become more popular in forecasting modeling because of its proficiency to handle the vanishing gradient problem in RNNs architecture. The combination of three gates with the cell state and hidden state helps the network to decide which moving values need to save, forget, remember, attention and the output format.

Gated Recurrent Unit (GRU) is another model which functions almost comparable with LSTMs but the architecture design is less complex compared LSTMs [39]. However, there are challenges occurred in performing RNN entity (LSTM & GRU) for sequential dataset that is training time is very slow due to backpropagation process.

In order to handle time series prediction problems, several attempts have been made in recent years to resolve issues using state-of-the-art algorithms. A standard Convolutional Neural Networks (CNN), developed specifically for image processing, have accomplished comparable and even exceeded LSTM on time series forecasting tasks [40].

Temporal Convolutional Neural Networks (TCNN) with 1D convolution Kernals which automatically learn time-translation invariant features from time series can be used in time series forecasting tasks. Based on updated CNN, forecasting time-series improved with its ability to extract long-term historical features by combining Casual Strategy and Dilated CNN, and by studying time-series data self-regressively and memorizing long-term historical information with a larger convolutionary receptive field [40–42]. This approach would therefore have a higher forecasting estimate by solving the long-term historical dependence problem.

In recent years, a variety of Ensemble Methods have been widely used to increase the ability of the model to generalize and improve additional accuracy [43] by about 2% in several real settings [44]. Hybrid Algorithms also have become popular, combining more than two algorithms that solve the problem, depending on the characteristics and structure of the data.

One of the research papers published is using the combination of Deep learning model Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) to investigate the performance of the network architecture to evaluate both the spatial and temporal evolution of the sea level [15]. The combination of the Convolutional Neural Network (CNN) and enhanced RNN model that is Long-Short Term Memory (LSTM) models are utilised to predict inter-annual sea level anomalies (SLAs) to capture both spatial and temporal relationships with higher accuracy. The combination of CNN and ConvLSTM networks has greatly increased accuracy

and forecasting has been consistent for a number of years. In addition, it addresses almost all spatial structures very well.

Another research paper is centered on the application of the Deep Neural Network (DNN) model to predict short-term future tide level using 10 years of historical dataset from 4 tide station [45]. The exact prediction of the tide level is important for near-shore marine activity and for anticipating sudden flash floods, particularly for people living on low-lying land. DNN model employed more hidden layers which executes different task. Significant parameter applied for tuning the DNN model in order to achieve higher accuracy are the activation function, learning rate ($\mu$), hidden layer architecture, dropout ratio and loss function. In order to improve performance, the DNN model is tuned to produce better performance using multiple hyperparameters.

Deep learning (DL) method viewed as a completed package to encounter the limitation of both traditional and shallow neural network models to forecast short and long term tidal changes. DL is capable to perform analysis on massive amount and complex characteristics of dataset within the time. In particularly, LSTM (with GRU) architectures designed to address sequential data or time series effectively [34]. The performance of individual Deep learning is evaluated based on the Mean Absolute Error Score (MAPE) and the Root Mean Square Error score (RMSE). Additionally, the optimization process helped to enhance the level of accuracy using multiple parameters such as the activation function, learning rate ($\mu$), hidden layer architecture, dropout ratio and loss function. It has been shown that the Deep Learning method can have higher predict accuracy over the short term and increase the forecast error with the forecasting days.

## 4 Conclusion and Future Perspectives

The implementation of various Deep learning methods under different conditions provides a diverse level of performance and the best solutions vary based on the characteristics of the problem domain. In the case of Tidal prediction, Irregular variations in tidal motions at high and low tides pose major challenges to learning algorithms. Apart from that, the consistency of the time series dataset use for future prediction also is very crucial. This means that handling of asynchronous data required an appropriate theoretical framework process. However, incongruent time-series configuration used in the same model will present different level of accuracy. In order to achieve a fully automated forecasting process, it is essential to adopt a forecasting framework that a similar time-series character between the problem domains.

**Acknowledgements** This work was partially supported by UMS Grant SDN0076.

## References


1. Devlin AT (2016) On the variability of Pacific Ocean tides at seasonal to decadal time scales: observed versus modelled
2. West BA, Gagnon IF, Wosnik M (2016) Tidal Energy Resource Assessment for McMurdo Station, Ant-arctica. Engineer Research and Development Center Hanover Nh Hanover United States
3. Peng D, Hill EM, Meltzner AJ, Switzer AD (2019) Tide gauge records show that the 18.61-year nodal tidal cycle can change high water levels by up to 30 cm. J Geophys Res: Oceans 124:736–749. https://doi.org/10.1029/2018jc014695
4. Wahl T, Brown S, Haigh ID, Nilsen JEØ (2018) Coastal sea levels, impact, and adaptation. J Mar Sci Eng 6:19. https://doi.org/10.3390/jmse6010019
5. Cazenave A (2018) Global sea-level budget 1993—present. Earth Syst Sci Data 10:1551–1590. https://doi.org/10.5194/essd-10-1551-2018
6. Herring SC, Hoerling MP, Kossin JP, Peterson TC, Stott PA (2015) Explaining extreme events of 2014 from a climate perspective. Bull Am Meteor Soc 96(12):S1–S172
7. Muis S, Verlaan M, Winsemius HC, Aerts JC, Ward PJ (2016) A global reanalysis of storm surges and extreme sea levels. Nat Commun 7(1):11969. https://doi.org/10.1038/ncomms11969
8. Nerem RS, Beckley BD, Fasullo JT, Hamlington BD, Masters D, Mitchum GT (2018) Climate-change–driven accelerated sea-level rise detected in the altimeter era. PNAS 115(9)
9. Dangendorf S, Marcos M, Wöppelmann G, Conrad CP, Frederikse T, Riva R (2017) Reassessment of 20th century global mean sea level rise. Research Institute for Water and Environment, University of Siegen, PNAS, June 6, 2017, 114(23)
10. Hay CC, Morrow E, Kopp RE, Mitrovica JX (2015) Probabilistic reanalysis of twentieth-century sea-level rise. Nature 517(7535):481–484
11. Church JA, White NJ (2011) Sea-level rise from the late 19th to the early 21st century. Surv Geophys 32(4–5):585–602. https://doi.org/10.1007/s10712-011-9119-1
12. Nerem RS, Chambers D, Choe C, Mitchum GT (2010) Estimating mean sea level change from the TOPEX and Jason altimeter missions. Mar Geod 33:435–446. https://doi.org/10.1080/01490419.2010.4910
13. Golledge NR (2019) Long-term projection of sea-level rise from ice-sheets. WIREs Clim Change 2020 11:e634. https://doi.org/10.1002/wcc.634
14. Simpson MJR, Ravndal OR, Sande H, Nilsen JEØ, Kierulf HP, Vestøl O, Steffen H (2017) Projected 21st century sea-level changes, observed sea level extremes, and sea level allowances for Norway. J Mar Sci Eng 2017(5):36
15. Braakmann-Folgmann A, Roscher R, Wenzel, Uebbing B, Kusche J (2017) Sea level anomaly prediction using recurrent neural net-works. Institute of Geodesy and Geoinformation, University of Bonn. arXiv:1710.07099v1 [cs.CV] 19 Oct 2017
16. S Ehsan, Begum RA, Md Nor NG, Maulud KNA (2019) Current and potential impacts of sea level rise in the coastal areas of Malaysia. In: IOP Conference Series: Earth and Environmental Science Paper, 228:012023. https://doi.org/10.1088/1755-1315/228/1/012023
17. Cai S, Liu L, Wang G (2018) Short-term tidal level prediction using normal time-frequency transform. Ocean Eng 156:489–499
18. Feng W, Zhong M, Xu HZ (2012) Sea level variations in the South China Sea inferred from satellite gravity, altimetry, and oceanographic data. Sci China Earth Sci. https://doi.org/10.1007/s11430-012-4394-3
19. Wahl T, Haigh ID, Nicholls RJ, Arns A, Dangen-dorf S, Hinkel J, Slangen A (2017) Understanding extreme sea levels for coastal impact and adaptation analysis. Nat Commun 8:16075
20. National Hydraulic Research Institute Malaysia (NAHRIM), (2010). The study of the Impact of Climate Change on Sea Level Rise on Malaysia Coastlines (Final Report) p172
21. Abubakar AG, Mahmud MR, Tang KKW, Hussaini A, Md Yusuf NH (2019) A review of modelling approaches on tidal analysis and prediction. In: The international archives of photogrammetry, remote sensing and spatial information science, vol XLII-4/W16

22. Breili K, Simpson MJR, Nilsen JEØ (2019) Observed sea-level changes along the Norwegian Coast. Mar Sci Eng 2017(5):29
23. Md Din AH, Ses S, Omar KM, Naeije M, Yaakob O, Pa' Suya MF (2014) Deprivation of sea level anomaly based on the best range and geophysical correction for Malaysian seas using radar altimeter database system (RADS). Jurnal Teknologi (Sciences & Engineering) 71(4):83–91
24. Fu Y, Zhou X, Zhou D, Sun W, Jiang C (2019) Sea level trend and variability in the South China Sea. ISPRS Ann Photogram, Remote Sens Spat Inf Sci IV-2/W5. https://doi.org/10.5194/isprs-annals-iv-2-w5-589-2019
25. Din AHM, Omar KM, Naeije M, Ses S (2012) Long-term sea level change in the Malaysian seas from multi-mission altimetry data. Int J Phys Sci 7(10):1694–1712. 2 March, 2012. https://doi.org/10.5897/ijps11.1596
26. Abdullah MH, Mahmud MR, Amat NA (2015) Variation of sea level and tidal behaviour during el-Nino/La-Nina: an example of Malaysian coastline 73(5):107–118. www.jurnaltek nologi.utm.my
27. Badejo OT, Akintoye SO (2017) High and low water prediction at Lagos Harbour, Nigeria. Niger J Technol. https://doi.org/10.4314/njt.v36i3.39
28. Srivastava PK, Islam T, Singh SK, Petropoulos GP, Gupta M, Di Q (2016) Forecasting Arabian Sea level rise using exponential smoothing state space models and ARIMA from TOPEX and Jason Satellite Radar Altimeter Data. Meteorol Appl 23:633–639
29. Hewamalage H, Bergmeir C, Bandara K (2019) Recurrent neural networks for time series forecasting: current status and future direction. Elsevier
30. Bandara K, Bergmeir C, Smyl S (2018) Forecasting across time series databases using recurrent neural networks on groups of similar series. arXiv:1710.03222
31. Liu Jiao, Shi Guoyou, Zhu Kaige (2019) High-precision combined tidal forecasting model. Algorithms 12:65. https://doi.org/10.3390/a12030065
32. Amuah VI, Boye CB (2018) Performance evaluation for mean sea level prediction using multivariate adaptive regression spline and artificial neural network. Ghana Min J 18(1):1–8
33. Hendri A, Suprayogi I, Zulfakar M, Ongko A (2017) Comparisons of tidal prediction analysis by using adaptive neuro fuzzy interference system (ANFIS) and artificial neural network (ANN). CSAI 2017, 5–7 Dec 2017, Jakarta, Indonesia. https://doi.org/10.1145/3168390.3168393
34. Le XH, Ho HV, Lee G, Jung S (2019) Application of long short-term memory (LSTM) neural network for flood forecasting. Water 11:1387. https://doi.org/10.3390/w11071387
35. Roshni T, Samui P, Drisya J (2019) Operational use of Machine Learning models for sea-level modeling. Indian J Mar Sci 48(09)
36. Imani M, Kao HC, Lan WH, Kuo CY (2018) Daily sea level prediction at Chiayi coast, Taiwan using ex-treme learning machine and relevance vector machine. Glob Planet Change 161:211–221
37. Lai G, Yang Y, Chang WC, Liu H (2018) Modelling Long- and short-term temporal patterns with deep neural networks. arXiv:1703.07015v3 [ cs.LG] 18 April 2018
38. Petnehàzi G (2019) Recurrent neural networks for time series forecasting. Doctoral School of Mathematical and Computational Sciences, University of Debrecen. arXiv:1901.000069v1
39. Yamak PT, Yujian L, Gadosey PK (2019) A com-parison between ARIMA, LSTM and GRU for time series forecasting. In: Proceedings of 2019 2nd international conference on algorithms
40. Geng Y, Su L, Jia Y, Han C (2018) Seismic events prediction using seep temporal convolution networks. J Electr Comput Eng 2019. Article ID 7343784
41. Van den Oord A, Dielemen S, Zen et al H (2016) WaveNet: a generative model for raw audio. In: Proceeding of the 9th ISCA speech synthesis workshop. Sunnyvale. CA, USA
42. Anastasia B, Sander B, Oosterlee CW (2017) Conditional time series forecasting with convolutional neural networks. In: Proceedings of the 26th international conference on artificial neural network (ICANN), Alghero, Italy

43. Wan R, Mei S, Wang J, Liu M, Yang F (2019) Multivariate temporal convolutional network: a deep neural networks approach for multivariate time series forecasting. Electronic 8:876. https://doi.org/10.3390/electronics8080876
44. Aggarwal CC (2018) Neural networks and deep learning: a textbook. Springer
45. Rasel RI, Uddin MN, Haroon A (2018) Application of deep neural network for predicting river tide level. In: International conference on innovations in sciences, engineering and technology (ICISET), https://doi.org/10.1109/iciset.2018.8745593

# The Most Potential Decision Tree Technique to Classify the Large Dataset of Students


**Afiqah Zahirah Zakaria, Ali Selamat, Hamido Fujita, and Ondrej Krejcar**

**Abstract** Education is one of the important fields in this challenging world. The researchers come out with the new perceptive, which is learning analytics that is a new invention for helping out the instructors, learners, and administrators. The use of learning analytics can be the medium for increasing the productivity of education for producing capable leaders in the future. Machine learning comes out with any type of techniques such as Decision Tree, Support Vector Machine, Naïve Bayes, and Ensemble Classifiers. However, both Decision Tree and Ensemble Classifiers are chosen as the best potential machine learning techniques to cope with the large database of students. The Boosted Tree of Ensemble Classifiers managed to get 99.6% accuracy of training 378,005 data of students regarding the Virtual Learning Environment (VLE).

**Keywords** Learning analytics · Machine learning · Decision tree · Big data · Ensemble classifiers

A. Zahirah Zakaria · A. Selamat (✉)
Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur, Malaysia
e-mail: aselamat@utm.my

A. Zahirah Zakaria
e-mail: afiqahzakaria08@gmail.com

A. Selamat
School of Computing, Faculty of Engineering, UTM and Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, Johor Bahru, Malaysia

H. Fujita
Faculty of Software and Information Science, Iwate Prefectural University, 152-52 Sugo, Takizawa 0200693, Iwate, Japan

A. Selamat · O. Krejcar
Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, Czech Republic



R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 724, https://doi.org/10.1007/978-981-33-4069-5_47

## 1 Introduction

The world is introduced to the new era that implemented data to all the things that happened in daily life. The main function of the data is that the data can be converted into useful information. The useful information needs to solve any problem and give a new perceptive for a certain situation.

One of the sectors that need to use large data is education and learning. This is due to both sectors always dealing with data every single time. The data is about many things, for example, the data about the institutions, the instructors or teachers, the students, the courses, and many more.

Learning Analytics (LA) is one of the formal analytic introduced for learning. Learning analytics is about the measurement, collection, analysis, and reporting of data about the learners and their context, for understanding and optimizing learning and the environments in which it occurs [1–5].

In learning analytics, it involves three main parties, which are administrators, instructors, and students. The administrators are the leader that gather all the information regarding the institutions, students, and instructors. While, the instructors collect the data about the courses, the examinations, the classes, the co-curriculum activities in many more.

The administrators recorded the students' information can be used as the input of analysis to measure the correlation between the information to gain the new informative result for the instructors. Learning analytics also become a subject particular concern especially considering the abundance a secondary data encompassing all aspects of students' trajectories across an academic curriculum [6]. Indeed, as technology progress, the understanding of online behavior is the main key to successful implementation in digital learning.

In this research, the authors focused on the effect of the number of splits towards Decision Tree techniques to increase the accuracy and cope with the large dataset. This is very important to find out the best techniques that can give the exact result of classifications.

This paper is coordinated as follows: Sect. 1 is an introduction about the information about the research including the field and techniques involved. Section 2 explained the previous implementation of current techniques in LA based on researches around the world. Section 3 explained the methodology involved in this research. Next, Sect. 4 discussed the results and discussion gathered in this research. Section 5 concluded this research and give opinions for future research.

## 2 Related Works

Both learning analytics (LA) algorithms and learning systems can be connected based on big data enable visualizing the integration of different data insights. Indeed, a set

of algorithms useful for the analysis in LA and also the massive data is pre-processing originally generated in the Massive Open Online Courses (MOOCs) [7].

MOOCs become famous since it starts in 2008. The traditional confined classroom education by universities also changing their practices by hosting MOOCs [8]. However, high dropout and failure rates become the main problems in MOOCs. The authors focused on investigating student dropout and the prediction results of machine learning models built based on different types of attributes and VLE [8]. The Gradient Boosting Machine proved the result prediction at 0.93 and high performance of AUC, up to 0.91 for dropout prediction based on students' interaction with VLE [8].

Other than that, by using the model of various approaches in LA and Educational Data Mining (EDM), the researchers got achievements in blended or online learning settings. Related to the methods, researchers used both statistical methods, logistic regression and probabilistic methods, Markov Chain, and Bayesian Classifier like Naïve Bayes. While Support Vector Machine (SVM) or decision trees like ID3 or C4.5 are machine learning for classification algorithms. The XGBoost showed the highest PR AUC value, which is 0.5652 [9].

The present grade of performance is explained in the emerging of education data mining and learning analytics promotes the acquisition of raw data for comprehensive visualization same as back-casting on learning behaviors [10, 11]. Both learning analytics and educational data mining are depending on two different machine learning and deep learning algorithms. According to the research, the target variable is 68.32% is classified correctly by the kNN algorithm. Besides, the student's performance on the final examination is evaluated correctly by Powers [10, 12]. Overall, the true positive rate of the models, 78.6% shows the effective result of decision tree algorithms as the study established [10].

The complements between a virtual learning environment and learning analytics paradigm that producing datasets for evaluating and recording the learning process of students and its both consideration and improvement in their various performance [13]. The implementation of a deep artificial neural network on a set of unique handcrafted features, and extracted from the clickstream data of virtual learning environment to forecast at-risk students contributing the estimations for early intervention [13]. The proposed model of deep neural network achieves a classification accuracy of 84–93% [13].

To determine the commitment development on student performance, machine learning algorithms are used to classify the low-engagement students in a social science course at the Open University (OU) [14]. The authors applied several machine-learning techniques to the dataset to anticipate the low-engagement students. The J48 has the highest accuracy for the training, which is 88.52% [14].

In the ML community, the ability to cope with the massive dataset is developed by ML algorithms become long-standing research before the advent of the "big data" era [15]. The Big Data dimension is divided into five, which are volume that related to the quantity or amount of data, velocity means the speed of data generation, variety related type, nature, and format of data, veracity is trustworthiness or quality of captured data and value about insight and impact.

By using the massive samples with high dimensionality, big data is characterized those conditions are prerequisites and imply the finding of datasets requires important computational and storage capabilities [16]. On the other hand, classification algorithms yield a good result in terms of large data processing that proved by the previous researchers [7].

Olson and Wu [17] suggested a classification model for unstructured documents via combining the Naïve Bayes classifier with the predictive analysis for learning systems [7, 17]. Besides, for the field of education EDM offers a range of algorithms [18]. Other than that, ML is one of the techniques of artificial intelligence. Due to that, ML algorithms can implement the complex pattern that is extracting the features from current data and allowed to make smart decisions regarding the existing data [19, 20].

There are collecting data, analyze the data, and provide appropriate suggestions and feedback to students to improve their learning [19, 20]. The predictive analytics can help the instructor to manage the students' activities with the learning material and the student's assessment scores are related to that student's engagement stage [14]. The computer abilities are increasing rapidly in e-learning, recommendation, pattern recognition, image processing, medical diagnosis, and many other domains using ML algorithms [21].

The sample data is used as inputs and tested with the new data in ML algorithms [22]. ML algorithms used by instructors that obtained student-related information in real-time that helps during early course stages [15, 23]. Both numerical and categorical forecaster variables with no doubt can build predictive models from student data and ML techniques [21]. The ML models could utilize big data techniques to achieve scalability in two categories, which are the basic middleware layer that reimplements existing learning tasks that can train in big data platforms and convert the single learning analytics into big data platforms [24].

## 3 Classifying the Large Dataset Using Decision Tree Techniques

This research is involving the large dataset about students' information based on student demographics, student activities, and module presentations regarding the VLE from Kaggle Dataset [25]. The final dataset is obtained by joining seven different tables. The Student Info table contains demographics details of students, Student Registration consists of information about the students registered or unregistered the courses. StudentVLE and VLE tables contain virtual learning environment information. Indeed, student assessment and assessment tables are about the assessments' information. Therefore, it involves 378,005 data including 26 predictors.

### *3.1 Data Pre-processing*

Data pre-processing is one of the techniques that can be implemented before the training session. Data pre-processing is important to prepare the data to get an accurate result during the training. It usually includes easy-to-measure variables selection, denoising, outlier detection, and missing values replacement is a crucial step process model design [26]. In this research, the authors mostly replace the missing data by zero to enhance the accurate result by the end of the training.

### *3.2 Decision Tree*

Decision Tree or DTs consists of simple decision rules inferred from the data features [27]. It likes a tree-like structure with internal nodes represented by rectangles and leaves represented by ovals. The internal nodes have two or more child nodes and it represents dataset characteristics; the branches represent the values of the characteristics. Besides, each leaf contains a class related to the dataset. It trained with a training set containing tuples to characterize the dataset with an unknown class label. After that, the dataset to process information in decision-making [21]. DT is a knowledge discovery process that included three phases, which are input, data analysis, and output.

DTs use a white box model, which is a noticeable situation in a model, the conditions are conveniently interpretable. Tree induction is a recursive top-down process that produces results in graphical form and/or rules expressions. The type of probability distributions by class and predictive variables does not require any prior assumptions [23]. It shows the attribute relationship nodes, branches, and children in the sense of the structure of a tree and association rules for natural language processing [28].

Therefore, in learning analytics, DT can visualize all the data regarding each parameter and link to each other is needed in the learning analytics process to overcome the problem faced by the students, instructors, and administrators. DTs algorithms proved it is the most viable method based on accuracy, predictive power, faster convergence, and can work with nonlinear and mixed predictor relationship datasets efficiently [23].

Other than that, decision trees are often used to construct trees and find predictive rules based on available data [18, 29]. DTs also good generalized and understandable classifiers [21]. They proved it when DTs can show the good visual rules to observe the influence of inputs variables and the correlations with the Final Grades [30].

### 3.3 Ensemble Classifiers

Ensemble Classifiers are focused on solving the imbalance data that usually happened in Big Data. Even though fault or abnormal data is happened for very small compared to normal data. This algorithm is the hybridization of two methods or techniques in machine learning or other related techniques to get the impressive result at the end of the classification training. It usually used to improve the execution time, accuracy, and overcome the problem caused by the data such as an imbalanced dataset. Besides, the increasing volume of data, computing resources may never be enough to analyze an entire big data set all at once. Based on this reason, the researchers are looking forward to finding out the best combination method for training the big data, especially for real-time data.

Boosted Tree is one of the classic well-performed ensemble algorithms that generate the final strong classifier by iterating weak classifiers through reweighting the samples in each iteration process [31, 32]. Several algorithms have been studied, but often boosting ensemble learning has yielded good results when solving the class imbalance problem in different domains such as fraud detection, medical diagnosis, and manufacturing quality control [33–35].

The Bagging scheme is an ensemble method that has been shown to have good performance in precise classification. Indeed, to give a final prediction, it combines the predictions of different training sets by a distinct model as the diversity is increased [36].

For the weak and unstable classifiers, the Bagging method proved that it can perform better than other techniques [36–40]. While RUSBoost or Random Under Sampling is an algorithm of an applicable unequal group of data [41].

### 3.4 Evaluation Parameters

The evaluation parameters are very significant for training the algorithm; hence, the most capable results can be revealing. The same type of evaluation parameters should be used so that, every technique can be comparing based on the parallel estimation.

In this research, there are three evaluation parameters are accuracy, prediction speed, and training time. Accuracy is the degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard.

The accuracy is a measure based on the formula below [42]:

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^{N} Ci, i}{\sum_{i=1}^{N} \sum_{j=1}^{N} C_{i,j}}$$

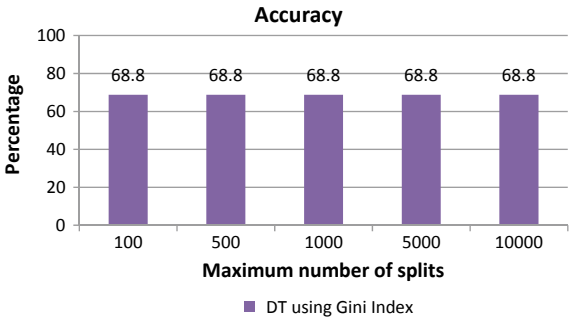



**Fig. 1** Accuracy of decision tree techniques based on gini diversity index

## 4 Results and Discussions

The authors used 378,005 of the dataset about the students' information related to the virtual learning environment via Kaggle Online Dataset [25]. The dataset will be going through Decision Tree and Ensemble Classifiers and be evaluated using accuracy as the evaluation parameters.

### *4.1 First Training Using Decision Tree Techniques*

At first, the research is done for Decision Tree by using Gini's diversity index as the split criterion based on the fixed learning rate = 0.1. Surprisingly, the result stated the same results of accuracy, which is 68.8% as shown in Fig. 1. Even though it was implemented five different types of a maximum number of splits, 100, 500, 1000, 5000, and 10,000 and cope with the large dataset.

### *4.2 Decision Tree Techniques*

Based on the result in Fig. 1, the authors tried an alternative to change the split criterion of the Decision Tree. Figure 2 shows the accuracy of the Decision Tree technique using the Twoing Rule with the same learning rate = 0.1. Besides, the training also is done using the three types of Boosted Tree, Bagged Tree, and RUSBoosted Tree. The training is done based on five different types of the maximum number of splits, which are 100, 500, 1000, 5000 and 10,000. All the ensemble classifiers are set up using the same number of learners, which are 100, 500, 1000, 5000, and 10,000.

The Boosted Tree shows the highest percentage of accuracy, 99.6% for the maximum number of splits, 10,000 compared to Decision Tree using Twoing Rule, Bagged Tree, and RUSBoosted Tree. The RUSBoosted Tree has the lowest accuracy at 40.2% for 100 maximum number of splits. Surprisingly, all the techniques are

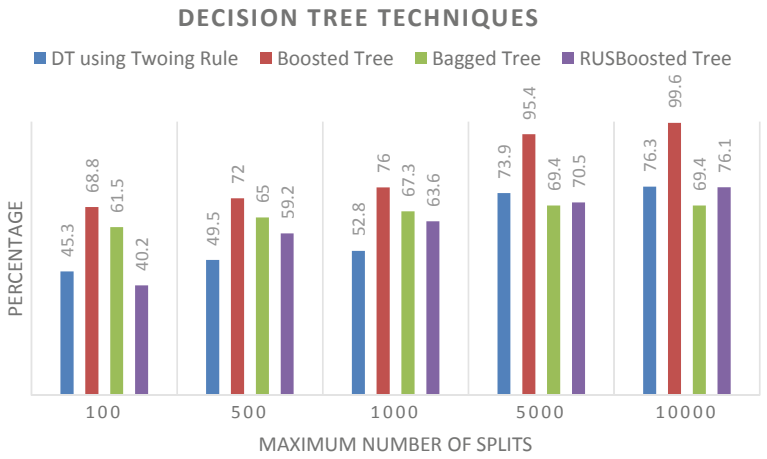



**Fig. 2** Accuracy of decision tree techniques

increasing towards a more accurate percentage as the maximum number of splits are increase even though the dataset is 378,005.

The Bagged Tree shows the small increase of each split. It starts from 61.5% of accuracy until the maximum accuracy is 69.4%. Besides that, RUSBoosted presents the difference of 35.9% between the highest and lowest accuracy. The highest is 76.1% of accuracy compared to 70.5% of accuracy for 5000 maximum number of splits.

Accuracy is important to measure the effectiveness of the techniques while undergoing the training phase. In this research, the combination of Decision Tree with the AdaBoost as Ensemble Classifiers with the increasing number of splits and the number of learners has the highest accuracy compared to other techniques. It shows that the highest number of splits and the number of learners can increase the accuracy of the technique. The number of splits means that the roots are divided widely. The increasing number of split helps training more effective and accurate results.

Accuracy is the first factor that is measured to choose the best classification technique. The accurate results can give the impact and solution for classification technique especially when it is related to the large dataset.

In this research, Boosted Tree can be seen as the best potential of a combination of Decision Tree with the other ensemble method. Even though it will take some time and slower speed compared to the others, but, it can be the best method as it can produce 99.6% accuracy of classification towards 378,005 datasets of VLE students. Therefore, it can be a huge opportunity for the Boosted Tree to be implemented for big data problems.

Besides, it proved that Boosted Tree can classify 99.6% accurately the student's information and the final examination results. The final examination results took based on the VLE tests that the students took earlier. Hence, it will help the instructors to find out the information of the fail students and monitor earlier to help the student

excellent in academics. Other than that, with accurate results, the administrator can identify the most effective tools that can be used for VLE. Therefore, without notice, the institution can have a better monitor VLE system and produce more excellent students in the future.

## 5 Conclusion

Learning analytics is one of the fields that need a lot of datasets to be trained and testing. This is due to the learning fields have many types of data to find out the solution and innovative for future education. With the growth of technology, machine learning is the opportunity for future inventions. However, machine learning contains a lot of techniques. The Boosted tree is proved as the most potential of the ensemble classifier based on Decision Tree and Adaboost. However, some improvement needs to be inventing into the technique to get the fastest speed and shortest time during the training period. The invention may be regarding the algorithm that can reduce the time and improve the speed as it needs to increase the number of splits and learners to get accurate accuracy. Therefore, it can be used widely in many fields to solve the problems regarding big data.

**Acknowledgements** The authors wish to thank Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876, and the Fundamental Research Grant Scheme (FRGS) Vot 5F073 supported under Ministry of Education Malaysia for the completion of the research. The works were also supported by the SPEV project, University of Hradec Kralove, FIM, Czech Republic (ID: 2102-2020). We are also grateful for the support of Ph.D. student Sebastien Mambou in consultations regarding application aspects.

## References

1. Siemens G, Long P (2011) Penetrating the fog: analytics in learning and education. Educause Rev 46(5):30
2. Gaševi D, Dawson S, Rogers T, Gasevic D (nd) Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success
3. Mothukuri UK, Reddy BV, Reddy PN, Gutti S, Mandula K, Parupalli R, Magesh E (2017) Improvisation of learning experience using learning analytics in eLearning of educational technology, vol 2. Sage Publications, Los Angeles, CA, pp 447–451
4. Siemens G (2013) Learning analytics: the emergence of a discipline. https://doi.org/10.1177/0002764213498851
5. Tempelaar D, Rienties B, Mittelmeier J, Nguyen Q (2018) Computers in human behavior student profiling in a dispositional learning analytics application using formative assessment. Comput Hum Behav 78:408–420. https://doi.org/10.1016/j.chb.2017.08.010
6. Ifenthaler D (2015). Learning analytics. In: Spector JM (ed) The sage encyclopedia
7. Hadioui A, Faddouli NEl, Touimi YB, Bennani S (2007) Machine learning-based on big data extraction of massive educational knowledge, 12(11):151–167

8. Jha NI, Ghergulescu I (2018) OULAD MOOC dropout and result prediction using ensemble, deep learning, and regression techniques
9. Hlosta M (2015) Ouroboros: arly identification of at-risk students without models based on legacy data
10. Alloghani M et al (2018) Application of machine learning on student data for the appraisal of academic performace. In: 2018 11th International Conference on Developments in eSystems Engineering (DeSE). IEEE, pp 157–162. https://doi.org/10.1109/DeSE.2018.00038
11. Algur SP, Bhat P, Kulkarni N (2016) Educational data mining: classification techniques for recruitment analysis. Int J Mod Educ Comput Sci 8(2):59–65
12. Desmarais MC et al (2012) A review of recent advances in learner and skill modeling in intelligent learning environments. User Model User Adap Inter 22(1–2):9–38. https://doi.org/10.1007/s11257-011-9106-8. https://search.proquest.com/docview/928407364?accountid=145382
13. Waheed H et al (2020) Computers in human behavior predicting the academic performance of students from VLE big data using deep learning models educational data, tools & technologies educational analytics. Comput Hum Behav 104(November 2018):106189. https://doi.org/10.1016/j.chb.2019.106189
14. Hussain M, Zhu W, Zhang W, Muhammad S, Abidi R (2018) Student engagement predictions in an e-Learning system and their impact on student course assessment scores
15. Zhou L, Pan S, Wang J, Vasilakos AV (2017) Neurocomputing machine learning on big data: opportunities and challenges. 237(September 2016):350–361.https://doi.org/10.1016/j.neucom.2017.01.026
16. Huck N (2019) Large data sets and machine learning: applications to statistical arbitrage. 278:330–342. https://doi.org/10.1016/j.ejor.2019.04.013
17. Olson DL, Wu D, (2017) Predictive models and big data. In: Predictive data mining models. Springer, pp 95–97. https://doi.org/10.1007/978-981-10-2543-3_8
18. Kuzilek J. Hlosta M, Herrmannova D, Zdrahal Z Wolff A (2015) OU analyze: analyzing at-risk students at the open university. In: Proceedings of first international workshop on visual aspects of learning analytics and knowledge conference (LAK 2015), Poughkeepsie, NY, USA, March 2015, pp 1–16
19. Costa LA (2019) Monitoring students performance in E-learning based on learning analytics and learning educational objectives, pp 9–10. https://doi.org/10.1007/s11704-015-4200-4
20. Paura L, Arhipova I (2014) Cause analysis of students' dropout rate in higher education study program. Procedia Soc Behav Sci 109:1282–1286
21. Kotsiantis S, Pierrakeas C, Pintelas P (2004) Predicting students performance in distance learning using machine learning techniques. Appl Artif Intell 18(5):411–426
22. Rizvi S, Rienties B, Ahmed S (2019) Computers & education the role of demographics in online learning; a decision tree-based approach. Comput Educ 137(April):32–47. https://doi.org/10.1016/j.compedu.2019.04.001
23. Kai S, Andres JML, Paquette L et al (2016) Predicting student retention from behavior in an online orientation course. In: Proceedings of 10th international conference on educational data mining, pp 250–255
24. Alloghani M, Al-jumeily D, Hussain A, Aljaaf AJ, Mustafina J, Petrov E (2018) Application of machine learning on student data for the appraisal of academic performace. In: 2018 11th International conference on developments in ESystems engineering (DeSE), pp 157–162. https://doi.org/10.1109/DeSE.2018.00038
25. Kuzilek J, Hlosta M, Zdrahal Z (2017) Data descriptor: open university learning analytics dataset, pp 1–8
26. Sliškovi D, Grbi R, Nyarko EK (2018) Data preprocessing in data-based process modeling. https://doi.org/10.3182/20090921-3-TR-3005.00096
27. Heuer H, Breiter A (2018) Student success prediction and the trade-off between big data and data minimization, pp 219–230
28. Mehedi M, Gumaei A, Alsanad A (2019) A hybrid deep learning model for efficient intrusion detection in a big data environment. Inf Sci. https://doi.org/10.1016/j.ins.2019.10.069

29. Buenaño-fern D, Gil D (2019) Application of machine learning in predicting performance for computer engineering students: a case study, pp 1–18
30. Qin SJ, Chiang LH (2019) Advances and opportunities in machine learning for process data analytics. Comput Chem Eng 126:465–473. https://doi.org/10.1016/j.compchemeng.2019.04.003
31. Fu Y et al (2019) Chemometrics and intelligent laboratory systems boosting classification tree-radial basis function network: application in metabonomics studies. Chemom Intell Lab Syst 193(July):103829. https://doi.org/10.1016/j.chemolab.2019.103829
32. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139
33. Obregon J, Kim A, Jung J (2019) RuleCOSI: combination and simplification of production rules from boosted decision trees for imbalanced classification. Expert Syst Appl 126:64–82. https://doi.org/10.1016/j.eswa.2019.02.012
34. Kim A, Oh K, Jung J-Y, Kim B (2018) Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles. Int J Comput Integr Manuf 31(8):701–717
35. Sun Y, Wong AK, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recogn Artif Intell 23(04):687–719
36. Moral-garcía S et al (2020) Bagging of credal decision trees for imprecise classification, 141. https://doi.org/10.1016/j.eswa.2019.112944.
37. Abellán J, Castellano JG (2017) A comparative study on base classifiers in ensemble methods for credit scoring. Expert Syst Appl 73:1–10. https://doi.org/10.1016/j.eswa.2016.12.020
38. Abellán J, Mantas CJ (2014) Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. Expert Syst Appl 41(8):3825–3830. https://doi.org/10.1016/j.eswa.2013.12.003
39. Abellán J, Masegosa A (2009) An experimental study about simple decision trees for bagging ensemble on datasets with classification noise. In: Symbolic and quantitative approaches to reasoning with uncertainty. Lecture Notes in Computer Science, vol 5590. Springer, pp 446–456. https://doi.org/10.1007/978-3-642-02906-6_39
40. Marqués A, García V, Sánchez JS (2012) Exploring the behavior of base classifiers in credit scoring ensembles. Expert Syst Appl 39(11):10244–10250. https://doi.org/10.1016/j.eswa.2012.02.092
41. Murugananthan V, Durairaj UK (2019) RUS boost tree ensemble classifiers for occupancy detection, (2):272–277. https://doi.org/10.35940/ijrte.B1048.0782S219
42. Abraham S, Huynh C, Vu H (2020) Classification of soils into hydrologic groups using machine learning. Data 5:2