جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

p-ISSN 2521-4209
e-ISSN 2521-4217

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.6 No.(2) December 2022

2022 2722

www.jst.uhd.edu.iq

Kurdistan Regional Government
University of Human Development

UHD Journal of Science and Technology

A periodic scientific journal issued by University of Human Development

Introduction

UHD Journal of Science and Technology (UHDJST) is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. UHDJST member of ROAD, e-ISSN: 2521-4217, p-ISSN: 2521-4209 and a member of Crossref, DOI: 10.21928/issn.2521-4217. UHDJST publishes original research in all areas of Science, Engineering, and Technology. UHDJST is a Peer-Reviewed Open Access journal with Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0). UHDJST provides immediate, worldwide, barrier-free access to the full text of research articles without requiring a subscription to the journal, and has article processing charge (APC). UHDJST applies the highest standards to everything it does and adopts APA citation/referencing style. UHDJST Section Policy includes three types of publications: Articles, Review Articles, and Letters.

By publishing with us, your research will get the coverage and attention it deserves. Open access and continuous online publication mean your work will be published swiftly, ready to be accessed by anyone, anywhere, at any time. Article Level Metrics allow you to follow the conversations your work has started.

UHDJST publishes works from extensive fields including, but not limited to:

- Pure Science
- Applied Science
- Medicine
- Engineering
- Technology

Scope and Focus

UHD Journal of Science and Technology (UHDJST) publishes original research in all areas of Science and Engineering. UHDJST is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. We believe that if your research is scientifically valid and technically sound then it deserves to be published and made accessible to the research community. UHDJST aims to provide a service to the international scientific community enhancing swap space to share, promote and disseminate the academic scientific production from research applied to Science, Engineering, and Technology.

SEARCHING FOR PLAGIARISM

We use plagiarism detection: detection; According to Oxford online dictionary, Plagiarism means: *The practice of taking someone else's work or ideas and passing them off as one's own.*

Section Policies

No.	Title	Peer Reviewed	Indexed	Open Submission
1	Articles: This is the main type of publication that UHDJST will produce	☑	☑	☑
2	Review Articles: Critical, constructive analysis of the literature in a specific field through summary, classification, analysis, comparison.	☑	☑	☑
3	Letters: Short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields.	☑	☑	☑

PEER REVIEW POLICIES

At UHDJST we are committed to prompt quality scientific work with local and global impacts. To maintain a high-quality publication, all submissions undergo a rigorous review process. Characteristics of the peer review process are as follows:

- The journal peer review process is a "double-blind peer review".
- Simultaneous submissions of the same manuscript to different journals will not be tolerated.
- Manuscripts with contents outside the scope will not be considered for review.
- Papers will be refereed by at least 2 experts as suggested by the editorial board.
- In addition, Editors will have the option of seeking additional reviews when needed. Authors will be informed when Editors decide further review is required.
- All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. Authors of papers that are not accepted are notified promptly.
- All submitted manuscripts are treated as confidential documents. We expect our Board of Reviewing Editors, Associate Editors and reviewers to treat manuscripts as confidential material as well.
- Editors, Associate Editors, and reviewers involved in the review process should disclose conflicts of interest resulting from direct competitive, collaborative, or other relationships with any of the authors, and remove oneself from cases in which such conflicts preclude an objective evaluation. Privileged information or ideas that are obtained through peer review must not be used for competitive gain.
- Our peer review process is confidential and the identities of reviewers cannot be revealed.

Note: UHDJST is a member of CrossRef and CrossRef services, e.g., CrossCheck. All manuscripts submitted will be checked for plagiarism (copying text or results from other sources) and self-plagiarism (duplicating substantial parts of authors' own published work without giving the appropriate references) using the CrossCheck database. Plagiarism is not tolerated.

For more information about CrossCheck/iThenticate, please visit
http://www.crossref.org/crosscheck.html.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. Open Access (OA) stands for unrestricted access and unrestricted reuse which means making research publications freely available online. It access ensures that your work reaches the widest possible audience and that your fellow researchers can use and share it easily. The mission of the UHDJST is to improve the culture of scientific publications by supporting bright minds in science and public engagement.

UHDJST's open access articles are published under a Creative Commons Attribution CC-BY-NC-ND 4.0 license. This license lets you retain copyright and others may not use the material for commercial purposes. Commercial use is one primarily intended for commercial advantage or monetary compensation. If others remix, transform or build upon the material, they may not distribute the modified material. The main output of research, in general, is new ideas and knowledge, which the UHDJST peer-review policy allows publishing as high-quality, peer-reviewed research articles. The UHDJST believes that maximizing the distribution of these publications - by providing free, online access - is the most effective way of ensuring that the research we fund can be accessed, read and built upon. In turn, this will foster a richer research culture and cultivate good research ethics as well. The UHDJST, therefore, supports unrestricted access to the published materials on its main website as a fundamental part of its mission and a global academic community benefit to be encouraged wherever possible.

Specifically:

- The University of Human Development supports the principles and objectives of Open Access and Open Science
- UHDJST expects authors of research papers, and manuscripts to maximize the opportunities to make their results available for free access on its final peer-reviewed paper
- All manuscript will be made open access online soon after final stage peer-review finalized.
- This policy will be effective from 17th May 2017 and will be reviewed during the first year of operation.
- Open Access route is available at http://journals.uhd.edu.iq/index.php/uhdjst for publishing and archiving all accepted papers,
- Specific details of how authors of research articles are required to comply with this policy can be found in the Guide to Authors.

ARCHIVING

This journal utilizes the LOCKSS and CLOCKSS systems to create a distributed archiving system among participating libraries and permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

LOCKSS: Open Journal Systems supports the LOCKSS (Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. LOCKSS is open source software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

CLOCKSS: Open Journal Systems also supports the CLOCKSS (Controlled Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. CLOCKSS is based upon the open-source LOCKSS software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

PUBLICATION ETHICS

Publication Ethics and Publication Malpractice Statement

The publication of an article in the peer-reviewed journal UHDJST is to support the standard and respected knowledge transfer network. Our publication ethics and publication malpractice statement is mainly based on the Code of Conduct and Best-Practice Guidelines for Journal Editors (Committee on Publication Ethics, 2011) that includes;

- General duties and responsibilities of editors.
- Relations with readers.
- Relations with the authors.
- Relations with editors.
- Relations with editorial board members.
- Relations with journal owners and publishers.
- Editorial and peer review processes.
- Protecting individual data.
- Encouraging ethical research (e.g. research involving humans or animals).
- Dealing with possible misconduct.
- Ensuring the integrity of the academic record.
- Intellectual property.
- Encouraging debate.
- Complaints.
- Conflicts of interest.

ANIMAL RESEARCHES

- For research conducted on regulated animals (which includes all live vertebrates and/or higher invertebrates), appropriate approval must have been obtained according to either international or local laws and regulations. Before conducting the research, approval must have been obtained from the relevant body (in most cases an Institutional Review Board, or Ethics Committee). The authors must provide an ethics statement as part of their Methods section detailing full information as to their approval (including the name of the granting organization, and the approval reference numbers). If an approval reference number is not provided, written approval must be provided as a confidential supplemental information file. Research on non-human primates is subject to specific guidelines from the Weather all (2006) report (The Use of Non-Human Primates in Research).
- For research conducted on non-regulated animals, a statement should be made as to why ethical approval was not required.
- Experimental animals should have been handled according to the highest standards dictated by the author's institution.
- We strongly encourage all authors to comply with the '*Animal Research: Reporting In Vivo Experiments*' (ARRIVE) guidelines, developed by NC3Rs.
- Articles should be specific in descriptions of the organism(s) used in the study. The description should indicate strain names when known.

ARTICLE PROCESSING CHARGES

UHDJST is an Open Access Journal (OAJ) and has article processing charges (APCs). The published articles can be downloaded freely without a barrier of admission.

Address

University of Human Development, Sulaymaniyah-Kurdistan Region/Iraq
PO Box: Sulaymaniyah 6/0778

Contact

Principal Contact

Dr. Aso Darwesh

Editor-in-Chief

University of Human Development – Sulaymaniyah, Iraq

Phone: +964 770 148 5879

Email: jst@uhd.edu.iq

Support Contact

UHD Technical Support

Phone: +964 770 247 3391

Email: jst@uhd.edu.iq

Contents

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Impact of Technological Burden on Knowledge Management Functions in Jordanian Industrial Companies

Muzhir Shaban Al-Ani[1], Shawqi N. Jawad[2], Suha Abdelal[2]

[1]Department of Information Technology, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq, [2]Department of Management, College of Business, Amman Arab University, Amman, Jordan

ABSTRACT

The goal of this study is to see how electronic information overload affects knowledge management functions in Jordanian businesses. All Jordanian industrial enterprises registered on the Amman Stock Exchange were included in the study's sample. Three hundred and seventy-three people were chosen at random from a simple random sample of 30% of the study population of 1242 senior and intermediate managers in the research community. Following the retrieval of the surveys, 206 questionnaires were found to be valid for analysis. It was used to do descriptive and heuristic statistical procedures, like simple and multiple regression analysis. The SPSS.16 application was used to do this. The study ends with the following findings: Electronic information overload (technological overload) has a statistically significant influence on knowledge management functions (acquisition, generation, transmission, exchange, and application) in Jordanian industrial companies. This work made a number of recommendations as a result of its findings, including: Adopting an organizational aspect that suits the nature of the tasks that industrial companies in Jordan perform, as well as providing technical capabilities to reduce the electronic information overload that these companies face while performing their tasks.

Index Terms: Knowledge Management, Organizational Overload, Statistical Analysis, Jordanian Industrial Companies

1. INTRODUCTION

These days, there has been a rapid acceleration of change toward the knowledge economy and the information economy. Knowledge is an essential ingredient for driving economic growth in countries [3]. Knowledge is already an intangible asset of the organization, leading organizations to reprioritize their efforts [49]. As a result, many technological applications that strengthened organizational capabilities and created a massive flow of information and their use in organizations were developed [55]. This led to the development of new trends in the management of organizations based on these ideas.

As a result of the information and technical revolution in all sectors of knowledge, today's business organizations confront several obstacles [10]. Therefore, senior management must be able to strengthen its role in investing in contemporary technology and expertise to improve its capacity to respond to the unpredictable environment and its demands [11].

As a result, the foundations of thinking and theoretical frameworks capable of fulfilling the organization's aims must be identified [12].

The rapid shift in the business environment has an impact on business organizations, particularly industrial enterprises [13].

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp1-10

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Muzhir Shaban Al-Ani, Department of Information Technology, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq. E-mail: muzhir.al-ani@uhd.edu.iq

Received: 21-01-2022 Accepted: 27-06-2022 Published: 01-07-2022

The factors that led to the burden of electronic information were how easy it was to get and store information in electronic databases, how often it was used, and how long it was kept [15].

Knowledge management is one of the modern topics in the field of management and business and is of great interest to those involved in business organizations [54]. This interest has also grown in the adaptation of various types of organizations to knowledge application [16]. Knowledge management is also important for the growth of current businesses and their ability to handle future problems [16]. Knowledge management's importance in corporate organizations lies not in the knowledge itself, but in the value, it adds to these firms. It also helps firms transition to a knowledge economy that prioritizes knowledge capital investment [6].

Due to the rapid technological development, the business environment of organizations is characterized by rapid change and is dominated by the ICT revolution [34]. Knowledge is the strength of organizations to ensure their growth and sustainability [32]. Knowledge is shared with participation and increased by practice and use [38]. Knowledge is an important resource that contributes to the success of different organizations and is governed by three fundamental characteristics, according to which knowledge is an important economic resource [42]. It is a leading sector of the contemporary economy and it can be traded indefinitely across one organization and among others in general [42].

Modern corporate organizations strive to adapt at every step of the knowledge economy's evolution to meet the demands of the time [48]. Electronic information systems now serve as the foundation for management and productivity operations in all types of businesses [58]. These systems are important for things such as business, marketing, and productivity [58].

1.1. Statement of the Problem

The organization and its employees are burdened with information that needs attention, research, and treatment. Businesses, particularly Jordanian industrial companies, deal with a vast volume of electronic data and information, which weakens their position in making various judgments and leads to mistakes due to the excessive weight placed on understanding information elements. This necessitates businesses to develop new and inventive ways to collaborate constructively to tackle these challenges, with the need to determine control techniques for the implementation of this mechanism to establish its possibilities.

"Knowledge Management Functions Used in Jordanian Industrial Companies: A Study of Technological Burden" is the goal of the study.

1.2. Research Questions

This research is implemented through addressing the following questions:

- According to managers, what kind of technology and what kind of technical skills do Jordanian industrial businesses have?
- What do managers think about the effects of technology load on the functions of knowledge management?
- Managers in Jordanian factories think about the technological load in two ways: How much technology there is and how well they can use it.

1.3. Research Objectives

The research aims to accomplish the following goals:

- Assessing the influence of electronic information (technological load) factors on knowledge management functions in the firms studied.
- Knowledge management functions can be good and bad. Find out how to improve them.
- Measuring the level of use of knowledge management functions by Jordanian industrial firms to come up with ideas for how to handle electronic data (technological burden).

1.4. Research Hypothesis

In Jordanian industrial organizations, there is no statistically significant influence of technological load (technology type and technical potential) on knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) at the level ($\alpha \leq 0.05$).

1.5. Research Model

The variables in the study model are chosen in accordance with the research's problem and hypothesis, as well as the study's purpose and particular goal (Fig. 1).

2. ELECTRONIC INFORMATION BURDEN

The electronic information burden in businesses companies is very important and needs a series of effective measures to overcome it.

Frank defined the burden of information as the point at which individuals' processing of information reaches its highest level, and therefore, the ability of individuals to process that information is reduced [52]. Bryant emphasizes

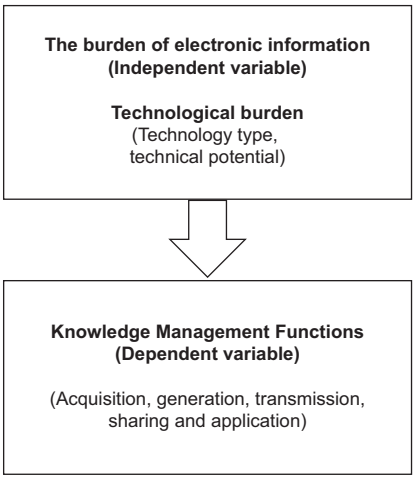


Fig. 1. Research model.

that the information burden occurs because there is more information than knowledge workers can absorb and determine what they need [28]. It has been shown that the information burden is attributed to the following elements: Multiple channels of information, time limiters, noise, and the volume of information coming in.

The burden of information is more pronounced in the fields of business in general and commercial business in particular. This confirms that the burden is a natural and inevitable condition and has several reasons for many of the developments and discoveries of the global era [21]. Several things show the burden of having too much information, like how people communicate, how they store and retrieve information, and how they make financial decisions.

Individuals are constantly exposed to a large amount of information that they obtain through their daily work, prompting them to refuse to receive this information and not allocate sufficient time to resolve the communication content. Choi *et al.* also showed that fear of dealing with information, and the inability to concentrate on memory-related problems may result in distracting thoughts and a lack of attention [22]. Such symptoms are reflected in the effectiveness of both the individual and the organization in their handling of information. According to Himma, the information burden arises from individuals' or the organization's management's frustration with not having access to the required information [37].

Filej *et al.* reported in his study that new information and communication technologies aim to facilitate rapid access to information [21]. Therefore, they cause a high overload of information, especially with push systems, which provide information to the user without any request for such information [50], [59]. Choi *et al.* explained that information technology plays an important role in accomplishing tasks, especially in business, and is an integral part of the manager's work [22]. Friedrich *et al.*, 2020 [31] reported that ICTs have increased access to information, processed it and produced new information, resulting in a burdensome information burden for managers [31]. Consider, but do not overlook, the role of information technology is assisting in the reduction of the burden through large-scale information processing. Mengis and Eppler emphasize that the development of technology has helped to increase the amount of information flowing to the stakeholders until it has become the main reason for generating the overload of information, directly or indirectly [27].

Below are a number of concepts that illustrate the definition of knowledge and related matters:
Knowledge: Knowledge is the product of data, intuition, and experience [61]. As a result, knowledge is information that has been organized, digested, and structured in such a way that it may be applied [36]. Organizational routines, rules, processes, documents, and practices involve a mix of contextual information, values, expertise, new information, and new expertise that exist in knowing minds, organizational routines, rules, processes, documents, and practices [4]. Knowledge is also recognized as a crucial component and a source of intellectual capital in today's enterprises, and it grows as a result of learning and practice [5]. As knowledge is a product of both the organization and the individual's practice, experience, judgments, and values, it is expressed in the process of applying knowledge to specific goals [8].

Existing two types of knowledge: There are two types of knowledge: Explicit and tacit. Explicit knowledge is information that can be shared across organizations, groups, and individuals, and it may be kept electronically, documented, conveyed, and used in a variety of ways, including knowledge maps. Tacit knowledge is knowledge that is stored in human thoughts and behavior and is difficult to record and transmit to others [1]. Tacit knowledge is knowledge that comes from past experiences and is hard to write down and pass on to others.

Knowledge management: Knowledge management is defined as "doing what it takes to maximize the value of knowledge resources" [30]. Because knowledge management is the gateway to adding and producing value by synthesizing

knowledge pieces to generate top knowledge combinations, the purpose of data, information, and knowledge will change [2], [5]. Knowledge management encompasses learning and adaptation, improving the creative process, sharing, and making the best use of these assets [40]. Knowledge management is defined as "successful learning processes linked with the exploitation, investigation, and sharing of human knowledge," according to the author (explicit and tacit). It improves performance and intellectual capital by utilizing proper technology, civilization, and culture.

Knowledge management contributes to the development of knowledge that attempts to enhance the success of companies through four dimensions, according to the organizational cooperation of knowledge management [30]. Processes, products, and people, as well as organizational performance, are all influenced by these characteristics.

Knowledge management functions can be divided into five functions as below:

- Knowledge acquisition: A function that attempts to collect and gain knowledge from a range of recorded sources as well as undocumented knowledge that is stored in people's thoughts and issued through their actions. Knowledge can be acquired from stakeholders and experts, with information technology playing a key role in data capture, classification, processing, and harnessing to generate a competitive advantage for the organization [25], [29], [41], [51], [56], [60], [62], [64]. It takes a lot of work to get new knowledge [41].
- Knowledge generation: It implies that information is created from a variety of sources and channels to expand organizational memory vaults and enable the company to discover innovative solutions to its issues, resulting in innovation. Individuals are involved in the creation of knowledge within the company. Social involvement, embodied external knowledge, integrated internal knowledge, and synthetic knowledge are the four ways that knowledge will be passed on to new people [60], [61], [62], [63], [64].
- Knowledge transfer: Leadership, absorptive capability, organizational structure support, degree of complexity, degree of privacy, and reliability of knowledge vocabulary are all elements to consider while transferring knowledge [16], [19], [23], [40], [44], [53].
- Knowledge sharing: Knowledge sharing is critical in modern businesses for production, adapting to external factors, outperforming rivals, promoting opportunities, and sustaining their efficacy. Individuals are sharing their expertise and experience formally through frequent formal meetings, which is expanding the organization's knowledge base. As well as people's involvement in knowledge with others, which avoids the loss of such information and its fading with time. Furthermore, sharing knowledge between organizations increases the amount of knowledge stored in organization warehouses [17], [20], [35], [39], [43], [45], [47], [56], [57], [63], [64].
- Knowledge application: The purpose of knowledge management is for modern enterprises to use knowledge retrieval techniques and web-based technology platforms to apply knowledge. Furthermore, these systems allow for the timely and appropriate access, use, and transfer of information, as well as communication with the appropriate individual [18], [33], [63], [64].

Tashkandi and Zakia examined the importance of knowledge management and the extent of its application in the management of education in the city of Mecca and concluded that the members of the study community recognize the importance of knowledge management and employment, but their management does not give priority to knowledge management [9]. The reason for this is that knowledge management is a modern area that organizations seek to adopt. Dubosson and Fragniere emphasized that the information burden affects the efficiency of organizations [26]. The burden is a curse and a real concern for the management of the organization, perhaps because the new IT trends are not fully absorbed, also showed that the information burden affects managers differently [22]. The role of the Director changed as he spent more time processing information and dealing with technology and less time managing staff, and he considered that the regulatory environment was the primary cause of the phenomenon of information burden, followed by technology and personal factors. Manovas concluded that the successful transfer of knowledge in an IT project must have a solid knowledge base and practical knowledge capabilities to ensure successful transfer of knowledge and that a culture of learning, sharing, collaboration technology, and incentive systems is important elements of the structure [46].

To measure the impact of organizational culture on knowledge management, the study of the convicted Rawluk *et al.* showed the impact of organizational culture factors individually and collectively in the management of knowledge as a whole, and its individual processes (knowledge generation, sharing, and application) [14]. Leadership was the most influential factor in organizational culture in implementing knowledge management. On the impact of knowledge management in achieving organizational creativity, Rawluk *et al.* [14]

showed that there is a clear awareness among employees of the need to adopt creative ideas in the technical fields to improve and develop. The organizational units in the research organizations want to adopt knowledge in all fields, through the adoption of expansion strategies in the scientific fields and the creation of new scientific departments.

3. METHODOLOGY

3.1. Research: Questionnaire Design

In the preceding sections, the electronic information overload in enterprises and their duties was discussed. This section will look at how to create the questionnaire that is necessary for this study. The questionnaire employs a five-field Likert scale, with strong agree, agree, neutral, disagree, and strongly disagree as the options. According to the study paradigm, the questionnaire is divided into two parts: An independent variable (technological burden) and a dependent variable (technological burden) (knowledge management functions).

- Independent variable (technological burden) (Fig. 2): This field is divided into two parts: Type of technology and technical potential and varied types of technology may not suit the type of task required, such as hardware, software, networks, and information systems, which are necessary to process data and information to accomplish various tasks. Human capabilities and infrastructure that support IT from databases and information processing systems, as well as specialists in data collection and analysis, maintenance workers, and equipment operators, are examples of technical potential.
- Dependent variable (knowledge management functions) (Fig. 3): Knowledge acquisition, knowledge generation, knowledge transfer, knowledge sharing, and knowledge application are the five components of this discipline. Knowledge acquisition refers to obtaining information from both internal (such as learning, cooperation, and employee feedback), as well as external (such as training programs and workshops and knowledge databases) sources (such as customers, consultants, competitors, competent personnel, attract experienced, and establish relationships with allies and partners). Knowledge generation is concerned with equipping workers in the knowledge field with graphical analysis, and this is done through learning, teaching, research, and development. Knowledge generation is concerned with creating and deriving new creative knowledge from existing knowledge through the organization to secure knowledge types for the benefit of future decisions, which are concerned with equipping workers in the knowledge field

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1	The company uses the right hardware and tools to obtain the appropriate information.					
2	The company uses various programs to perform various electronic operations.					
3	The company provides systems to operate networks electronically.					
4	The company keeps updating the software used for its operations.					
5	Information technology enhances the company's ability to filter information.					
6	The company employs specialized software designers in proportion to the technological burden.					
7	The company employs specialized equipment operators.					
8	The company uses multiple databases supporting the role of IT operations.					
9	The company has an infrastructure to support information technology, such as information and communication networks, servers, peripherals and accessories.					
10	The company employs data collection specialists.					
11	The company provides feasible information processing systems.					
12	The company has a specialized IT maintenance unit.					
13	The company employs professionals who specialize in data processing and preparation for use.					

Fig. 2. Technological burden questioner.

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1	The company receives feedback from employees on a permanent basis.					
2	The company uses training programs and workshops as a way to equip employees with the necessary knowledge.					
3	The company acquires knowledge from its partners or allies by establishing relationships with them.					
4	The company is keen to provide employees with information consistent with its products					
5	The company recruits experienced and competent staff to work for it to enhance its knowledge.					
6	The company always synthesizes information collected from multiple sources, in order to generate new knowledge.					
7	The company adopts advanced R&D policies to generate new knowledge.					
8	The company provides incentives for new innovations and knowledge.					
9	The company encourages brainstorming among employees to generate new ideas.					
10	The company seeks to meet its knowledge needs by bridging the knowledge gap.					
11	Company information flows smoothly across functional boundaries.					
12	The company contributes in sending scholarships for specializations in order to transfer knowledge.					
13	IT helps bring people in need of knowledge closer to those who have it.					
14	The company encourages dialogue between employees to impart knowledge.					
15	The company makes periodic transfers between departments and departments as a means of knowledge sharing.					
16	The company employs informal meetings and dialogues for the purposes of expanding sharing of knowledge.					
17	The company has an atmosphere of mutual cooperation to support knowledge sharing.					
18	The company fosters a culture of knowledge sharing among employees.					
19	The company provides multiple channels for knowledge sharing (Internet, Extranet and Intranet).					
20	The company holds meetings to discuss its annual reports to get feedback.					
21	The organization holds training courses on how to use and apply the knowledge gained to achieve specific objectives.					
22	Directors recognize that the Organization has a non-invested knowledge balance.					
23	The company uses modern technologies to apply knowledge and invest its returns.					
24	The management of the company is keen to use the new knowledge generated by the company.					
25	The company is keen to ensure that employees are aware of the methods of applying the acquired knowledge.					

Fig. 3. Knowledge management functions questioner.

with graphical analysis, and this is done through learning, teaching, research, and development. Knowledge transfer refers to the proper conveyance of specific knowledge to a specific individual (communications, reports, bulletins, staff movements, and use of technical tools to enhance knowledge transfer) in a formal and informal way (individual informal meetings) at the correct cost and at the right time.

Individuals share and circulate numerous sorts of knowledge in the context of knowledge sharing. Securing collective collaboration among them, interacting with others' conversations both within and outside the company, reaching out and working on the same document concurrently from many locations to develop fresh creative ideas everywhere in an organization, knowledge application refers to how people and resources are developed, how businesses are made better, how technology is used to get information, and how people can share information about what they know.

3.2. Research: Population and Sample

Industrial enterprises registered on the Amman Stock Exchange make up the study's population. All managers in senior management (general managers, their assistants, or their representatives) as well as managers in middle management were included in the sample and analysis unit (directors of the main departments and heads of departments). The sampling and analysis unit has a population of 1242 people. A random selection of 30% of them was chosen to form the sample (373 people). The questionnaires were then sent out to that group. There were 206 questionnaires that could be used for statistical analysis, accounting for 55% of the total distributed questionnaires.

Secondary sources, which contain data and material published in various library sources, have been permitted for assessment of the literature and prior works to meet the study's goal. In this case, the questionnaire was the main source of data. It was used to manage the questions and test the hypotheses in the applied part of a study.

4. RESULTS AND ANALYSIS

The relative importance was determined in the respondents' perceptions of the study questions based on the Likert five-point scale. A low (<2.33), middle (2.33–3.66), and high (3.77 and above) are used as shown in Fig. 4.

Question 1: How do managers working in Jordanian industrial enterprises perceive the technological load (kind of Fig. 1 shows that the sample's perceptions of the technological burden were high in both dimensions of the type of technology and technical possibilities, with the general arithmetic mean of the type of technology (4.20) and the standard deviation of the standard deviation of the standard deviation of (0.55).

This is because businesses recognize the significance of the technology they employ as the foundation for data and information filtering. It is very important for businesses to have electronic network operating systems because they have to deal with a lot more information.

Question 2: What are the perspectives of managers in Jordanian industrial businesses that operate in knowledge management functions (acquisition, generation, transfer, sharing, and application of information)?

Fig. 1 shows that the respondents' level of response was high, with an arithmetic mean of (3.71) and a standard deviation of (0.57). While the high degree of knowledge sharing revealed a significant influence of the electronic burden on knowledge sharing, the results also revealed a significant impact of the electronic burden on knowledge sharing. The researchers argue that the availability of technological devices and equipment hinders the establishment of a favorable regulatory environment, beginning with the speed with which information is delivered across numerous communication networks both internally and externally. Finally, respondents gave a positive response to the application of knowledge.

According to the researchers, this is due to the capacity of the investigated firms to apply current techniques in the application of information and to strive to communicate it to all people, as well as the use of fresh knowledge that is relevant and produces excellent results.

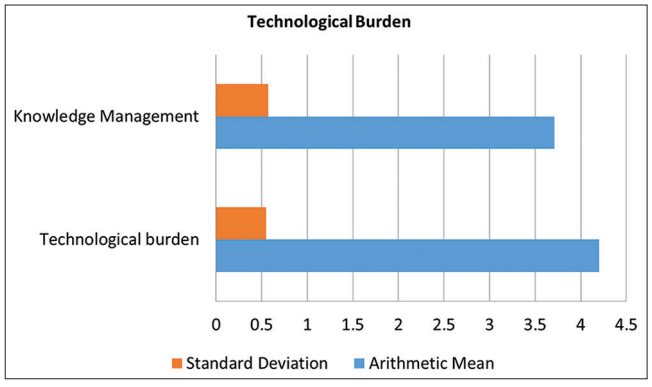


Fig. 4. Arithmetic mean and standard deviation (Technological Burden).

In this section, the hypotheses of the study are tested, where simple regression, multiple regression, and other tests are used to validate the hypotheses. These tests are the F-test for the significance of the regression model, the *t*-test for the effect of significance, and the value of R^2 (coefficient of determination), to find out the percentage interpreted by the independent variables in the dependent variable, depending on the statistical significance values extracted using the statistical software as below:

The hypothesis of the study: In Jordanian industrial organizations, there is no statistically significant influence of technological load (type of technology and technical skills) on knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) at the level of (0.05).

Table 1 findings show that technical load is the most important variable that determines knowledge management

TABLE 1: The influence of technological strain on the dimensions of knowledge management functions

Dependent variable	Standard deviation	Coefficient of determination	F-value	Degree of freedom	Significance level	Regression coefficients				
						Independent variable	β	Standard error	t-test	Significance level
Acquisition of Knowledge	0.515	0.266	73.790	1.204	0.000	Technological Burden	0.510	0.059	8.590	0.000
Generation of Knowledge	0.535	0.286	81.709	1.204	0.000	Technological Burden	0.572	0.063	9.039	0.000
Transmission of Knowledge	0.601	0.361	115.331	1.204	0.000	Technological Burden	0.678	0.063	10.739	0.000
Sharing of Knowledge	0.535	0.286	81.733	1.204	0.000	Technological Burden	0.577	0.064	9.041	0.000
Application of Knowledge	0.498	0.248	67.182	1.204	0.000	Technological Burden	0.569	0.069	8.196	0.000
Functions of Knowledge Management	0.618	0.382	126.324	1.204	0.000	Technological Burden	0.581	0.052	11.239	0.000

TABLE 2: The influence of technological load characteristics on knowledge management function dimensions

Dependent variable	Coefficient of determination	Standard deviation	F-test	Degree of freedom	Significance level	Regression coefficients				
						Independent variable	β	Standard error	T-Test	Significance level
Acquisition of knowledge	0.540	0.292	41.850	2.203	0.000	Type of technology	0.091	0.079	0.901	0.368
						Technical capabilities	0.445	0.480	5.495	0.000
Generation of knowledge	0.610	0.372	60.137	2.203	0.000	Type of technology	0.077	0.101	0.769	0.443
						Technical capabilities	0.549	0.080	6.835	0.000
Transmission of knowledge	0.536	0.287	40.825	2.203	0.000	Type of technology	0.269	0.103	2.625	0.009
						Technical capabilities	0.321	0.082	3.916	0.000
Sharing of knowledge	0.498	0.248	33.464	2.203	0.000	Type of technology	0.193	0.112	1.729	0.085
						Technical capabilities	0.369	0.089	4.136	0.000
Application of knowledge	0.518	0.268	37.137	2.203	0.000	Type of technology	0.269	0.095	2.821	0.005
						Technical capabilities	0.261	0.076	3.431	0.001
Functions of knowledge management	0.619	0.383	63.088	2.203	0.000	Type of technology	0.180	0.083	2.166	0.031
						Technical capabilities	0.389	0.066	5.861	0.000

functions. All models of progressive multiple regression are consistent with statistical tests, including F values and the *t*-test, implying that the initial zero hypothesis is rejected and the alternative hypothesis is accepted.

The influence of technological strain on the dimensions of knowledge management functions is shown in Table 1.

Table 2 shows that the multiple regression model used to assess the impact of the dimensions of the independent variable technological burden (type of technology and technical capabilities) on the variable of knowledge management functions is significant, and that the two variables together ($R^2 = 38.3\%$) account for the majority of the differences in knowledge management function values.

Table 1 shows the significance of the multiple regression model when it comes to the impact of type of technology and technical abilities on the dimensions of knowledge management functions. Technical capabilities and type of technology have a big impact on the two dimensions of knowledge transfer and application, but not so much on the technical sub-dimensions of knowledge acquisition.

5. CONCLUSIONS

In light of the prior talks, the study came to the following conclusions:

- The high importance of the technological burden in terms of the type of technology in Jordanian industrial companies, particularly in terms of the companies' use of appropriate devices and tools to obtain the required and appropriate information in their use through their reliance on activating local and spider networks.
- The findings of this study agreed with those of the Ching-Chiao *et al.* [24] study, which found that technology produces an information burden and that the sort of technology used is critical in decreasing that burden. The present study's findings also agreed with those of Dubosson and Fragniere [26] in that information technology is an information burden that reduces company efficiency.
- The study found that the technological burden is becoming more important in terms of technical skills in Jordanian industrial enterprises, particularly in terms of their capacity to have a supporting infrastructure for information technology, networks, servers, and other peripherals. This study supported the conclusions of the Choi *et al.* [22] study, especially in terms of the human component, where the larger the information load, the more time managers spend on analysis and audits, and the less time they spend on staff management. In the technological section of the study, the previous study Choi *et al.* [22] varied from the current one since technology was the second source of the information load phenomena.
- The study revealed that information generation piques people's attention. This is due to managers' capacity to diversify information sources, their focus on research and development, and an effort to bridge knowledge gaps as a result of changes in the workplace, as well as attention to the organizational and technical dimensions of knowledge creation. This result matched that of research on knowledge sources, acquisition, and transmission [9]. The findings are also in line with research Suhaimi [7] that looked at the characteristics and processes of knowledge management (acquisition, generation, transmission, distribution, and application).

6. RECOMMENDATIONS

As a result of their research, the researchers make a number of recommendations for Jordanian industrial companies to follow. These recommendations are: Adopting the type of technology appropriate to the environment in which Jordanian industrial companies operate in such a way as to reduce the burden of information in terms of the technological burden that may be exposed to them in carrying out their decision-making tasks.

- Jordanian industrial businesses should undertake a strategic study of the company's strengths and weaknesses as reflected in the performance of their departments and departments and decide their influence on raising or lowering the technical load in those departments and departments.
- Building organizations in their regulatory environments to make communication systems that work well, based on the idea of cutting down on technology so that they can reorganize their systems to be more effective.
- Organizations are anxious to guarantee that employees are aware of how to use previously acquired information and how to apply new knowledge developed by these companies based on their unique characteristics. Recognize that managers have a knowledge asset that has yet to be invested in and support this trend by offering training courses on how to use and apply that expertise to achieve certain goals.

REFERENCES

[1] A. Jazar and A. Talaat. "*Proposed Project for Knowledge Management in Jordanian Public Universities*". Unpublished Doctoral Thesis, Faculty of Higher Education Studies. Amman Jordan, Amman Arab University for Graduate Studies, 2005.

[2] M. Bataineh and Z. Mashaqbeh. "*Knowledge Management between Theory and Practice*". Jalis Al-Zaman Publishing House, Amman, 2010.

[3] S. N. Jawad, M. S. Al-Ani, H. A. Hijazi and H. Irshaid. "*Small Business Management, A Technology Entrepreneurial Perspective*". Safa Publishing House, Amman, Jordan, 2010.

[4] H. Hijazi. "*Measuring the Impact of Knowledge Management Perception on Employment in Jordanian Organizations: A Comparative Analytical Study between the Public and Private Sectors towards Building a Model for Knowledge Management Employment*". Unpublished Doctoral Thesis, Faculty of Administrative and Financial Studies, Amman Arab University for Graduate Studies, Amman, Jordan, 2005.

[5] N. Hamidi. "*Management Information Systems: Contemporary Entrance*". Wael Publishing House, Amman, Jordan, 2005.

[6] M. Ziadat. "*Contemporary Trends in Knowledge Management*". Safaa Publishing and Distribution House, Amman, Jordan, 2008.

[7] Z. Suhaimi. "*The Readiness of Public Organizations for Knowledge Management: A Case Study of King Abdul Aziz University in Jeddah*". An Introduction to the International Conference on Administrative Development: Towards Distinguished Performance of the Government Sector, Riyadh, 2009.

[8] S. Al. Sharfa. "*The Role of Knowledge Management and Information Technology in Achieving Competitive Advantages in Banks Operating in Gaza Strip*". Master of Business Administration, Islamic University, Gaza, 2008.

[9] Z. Tashkandi. "*Knowledge Management: The Importance and Extent of Application of its Operations from the Point of View of the Supervisors and Administrator's Departments of the Department of Education in Makkah and Jeddah*". Master Thesis, Umm Al-Qura University, 2009.

[10] A. Taher and I. Al-Mansour. "*Requirements for Sharing Knowledge and Obstacles Facing its Application in Jordanian Telecommunication Companies*". Presented to the Scientific Conference, Applied Science University, Amman, Jordan, 2007.

[11] M. S. Al-Ani and S. N. Jawad. "*Management Process and Information Technology*". Al-Ethaa Publishing House, Amman, Jordan, 2008.

[12] M. S. Al-Ani and S. N. Jawad. "*Business Intelligence and Information Technology*". Safa Publishing House, Amman, Jordan, 2012.

[13] A. A. Eniola, G. K. Olorunleke, O. O. Akintimehin, J. D. Ojeka and B. Oyetunji. "The impact of organizational culture on total quality management in SMEs in Nigeria". *Heliyon*, vol. 5, no. 8, p. e02293, 2019.

[14] A. Rawluk, R. M. Ford, L. Little, S. Draper and K. J. H. Williams. "Applying social research: How research knowledge is shaped and changed for use in a bushfire management organization". *Environmental Science and Policy*, vol. 106, pp. 201-209, 2020.

[15] A. R. Said, H. Abdullah, J. Uli and Z. A. Mohamed. "Relationship between organizational characteristics and information security knowledge management implementation". *Procedia Social and Behavioral Sciences*, vol. 12320, pp. 433-443, 2014.

[16] A. M. Abubakar, H. Elrehail, M. A. Alatailat and A. Elçi. "Knowledge management, decision-making style and organizational performance". *Journal of Innovation and Knowledge*, vol. 4, no. 2, pp. 104-114, 2019.

[17] S. Almahamid, A. Awwad and M. McAdams. "Effects of organizational agility and knowledge sharing on competitive advantage: An empirical study in Jordan". *International Journal of Management*, vol. 27, no. 3, pp. 387-404, 2010.

[18] M. Ariffin, N. Arshad, A. R. S. Shaarani and S. U. Shah. "Implementing knowledge transfer solution through web-based help desk system". *World Academy of Science Engineering and Technology*, vol. 21, pp. 78-82, 2007.

[19] E. Awad and H. Ghaziri. "*Knowledge Management*". Pearson Education Inc., Prentice Hall, United States, 2004.

[20] K. Bartol and A. Srivastava. "Encouraging knowledge sharing: The role of organizational reward systems". *Journal of Leadership and Organizational Studies*, vol. 9, no. 1, pp. 64-76, 2002.

[21] B. Choi, S. K. Poon and J. G. Davis. "Effects of knowledge management strategy on organizational performance: A complementarity theory-based approach". *Omega*, vol. 36, no. 2, pp. 235-251, 2008.

[22] B. Filej, B. Skela-Savič, V. H. Vicic and N. Hudorovic. "Necessary organizational changes according to Burke–Litwin model in the head nurses system of management in healthcare and social welfare institutions-The Slovenia experience". *Health Policy*, vol. 90, no. 2-3, pp. 166-174, 2009.

[23] J. Bou-Liusar and M. Segarra-Cipres. "Strategic knowledge transfer and its implications for competitive advantage: An integrative conceptual framework". *Journal of Knowledge Management*, vol. 10, no. 4, pp. 100-112, 2006.

[24] Y. Ching-Chiao, P. B. Marlow and C. S. Lu. "Knowledge management enablers in liner shipping". *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no. 6. pp. 893-903, 2009.

[25] W. Cohen and D. Levinthal. "Absorptive capacity: A new perspective on learning and innovation". *Administrative Science Quarterly*, vol. 35, no. 1, pp. 128-152, 1990.

[26] M. Dubosson and E. Fragniere. "The consequences of information overload in knowledge based service economies: An empirical research conducted in Geneva". *Service Science*, vol. 1. no. 1, pp. 56-62, 2009.

[27] M. Eppler and J. Mengis. "*A Framework for Information Overload Research in Organizations: Insights from Organization Science, Accounting, Marketing, MIS, and Related Disciplines*". ICA Working Paper, University of Lugano, Lugano, 2003.

[28] B. Furlow. "Information overload and unsustainable workloads in the era of electronic health records". The *Lancet Respiratory Medicine*, vol. 8, no. 3. pp. 243-244, 2020.

[29] J. Feliciano. "*The Success Criteria for Implementing Knowledge Management Systems in an Organization*". Doctoral Dissertation, Pace University, USA, 2006.

[30] A. Ferraris, C. Giachino, F. Ciampi and J. Couturier. "R&D internationalization in medium-sized firms: The moderating role of knowledge management in enhancing innovation performances". *Journal of Business Research*, vol. 128, pp. 711-718, 2019.

[31] J. Friedrich, M. Becker, F. Kramer, M. Wirth and M. Schneider. "Incentive design and gamification for knowledge management". *Journal of Business Research*, vol. 106, pp. 341-352, 2020.

[32] S. Goh. "Managing effective knowledge transfer: An integrative framework and some practice implications". *Journal of Knowledge*

Management, vol. 6, no. 1, pp. 23-30, 2002.

[33] R. Grant and C. Baden-Fuller. "A knowledge accessing theory of strategic alliances". *Journal of Management Studies*, vol. 41, no. 1, pp. 61-84, 2004.

[34] M. L. Grise and B. Gallupe. "Information overload: Addressing the productivity paradox in face-to-face electronic meetings". *Journal of Management Information Systems*, vol. 16, no. 3, pp. 157-186, 2000.

[35] D. Gurteen. "*Creating a Knowledge Sharing Culture*". Vol. 2. Knowledge Management Magazine, 1999.

[36] H. Biemans and C. Siderius. "Advances in global hydrology-crop modelling to support the UN's sustainable development goals in South Asia". *Current Opinion in Environmental Sustainability*, vol. 40, pp. 108-116, 2019.

[37] K. Himma. "A preliminary step in understanding the nature of a harmful information-related condition: An analysis of the concept of information overload". *Ethics and Information Technology*, vol. 9, no. 4. pp. 4, 2007.

[38] J. Hodge. "*Examining Knowledge Management Capability: Verifying Knowledge Process Factors and Areas in an Educational Organization*", Doctoral Dissertation, Northcentral University, 2010.

[39] M. Ismail and Z. Yusof. "The impact of individual factors on knowledge sharing quality". *Journal of Organizational Knowledge Management*, vol. 2010, pp. 13, 2010.

[40] A. Jashapara. "*Knowledge Management an Integrated Approach*". Pearson Education, Prentice-Hall, Hoboken, 2004.

[41] K. Mellahi and D. G. Collings. "The barriers to effective global talent management: The example of corporate élites in MNEs". *Journal of World Business*, vol. 45, no. 2, pp. 143-149, 2010.

[42] Y. L. Kim and W. Van Biesen. "Fluid overload in peritoneal dialysis patients". *Seminars in Nephrology*, vol. 37, no. 1, pp. 43-53, 2017.

[43] N. Leung and S. Kang. "Ontology-based collaborative Inter-organizational knowledge management network". *Interdisciplinary Journal of Information Knowledge and Management*, vol. 4. p. 699. 2009.

[44] L. Lin, X. Geng and A. Whinston. "A sender-receiver framework for knowledge transfer". *MIS Quarterly*, vol. 29, no. 2, pp. 197-219, 2005.

[45] K. Mahesh and J. Suresh. "Knowledge criteria for organization design". *Journal of Knowledge Management*, vol. 13, no. 4, pp. 41-51, 2009.

[46] M. Manovas, "Investigating the Relationship between Knowledge Management Capability and Knowledge Transfer Success". Mastery Degree, Concordia University, Canada, 2004.

[47] M. Mohayidin, N. Azirawani, M. Kamaruddin and M. Idawati. "The Application of knowledge management in enhancing the performance of Malaysian Universities". *Journal of Knowledge Management*, vol. 5, no. 3, pp. 301-312, 2007.

[48] G. B. Mulder. Management, husbandry, and colony health. In: *The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents*. Ch. 28. Academic Press, Cambridge, pp. 765-777, 2012.

[49] G. B. Mulder. "Perception as information processing". *Urban Ecology*, vol. 4, no. 2, pp. 103-118, 1979.

[50] M. Raoufi. "*Avoiding Information Overload-A study on Individual's Use of Communication Tools*". Proceeding of the 36th Hawaii International Conference on System Sciences, 2003.

[51] E. Reiter, A. Cawsey, L. Osman and Y. Roff. "*Knowledge Acquisition for Content Selection*". In: Proceedings of the Sixth European Workshop on Natural Language Generation, 1997, pp. 117-126.

[52] F. Ruff. "The advanced role of corporate foresight in innovation and strategic management-reflections on practical experiences from the automotive industry". *Technological Forecasting and Social Change*, vol. 101, pp. 37-48, 2015.

[53] W. Seidman and M. McCauley. "*Optimizing Knowledge Transfer and Use*". Cerebyte, Inc., Lake Oswego, Oregon, 2005.

[54] N. K. Sekaran and G. B. Seymann. "Hospital-based quality improvement initiatives". *Hospital Medicine Clinics*, vol. 3, no. 3, pp. e441-e456, 2014.

[55] J. Song, H. Zhan, J. Yu, Q. Zhang and Y. Wu. "Enterprise knowledge recommendation approach based on context-aware of time-sequence relationship". *Procedia Computer Science*, vol. 107, pp. 285-290, 2017.

[56] A. Tiwana. "*The Knowledge Management Toolkit: Orchestrating IT, Strategy and Knowledge Platform*". 2nd ed. Prentice Hall, Upper Saddle River, 2002.

[57] S. Wang. "*To Share or not to Share: An Examination of the Determinants of Sharing Knowledge via Knowledge Management Systems*". Doctoral Dissertation, Ohio State University, United States, 2005.

[58] E. Whelan and R. Teigland. "Transactive memory systems as a collective filter for mitigating information overload in digitally enabled organizational groups". *Information and Organization*, vol. 23, no. 3, pp. 177-197, 2013.

[59] T. Wilson. "Information overload: Implications for healthcare services". *Health Informatics Journal*, vol. 7, no. 2, pp. 112-117, 2001.

[60] R. Wong and T. Tiainen. "Are you ready for right knowledge management strategy: Identifying the potential restrains using the action space approach". *Frontiers of E-Business Research*, pp. 480-490, 2004.

[61] X. Xie, H. Zou and G. Qi. "Knowledge absorptive capacity and innovation performance in high-tech companies: A multi-mediating analysis". *Journal of Business Research*, vol. 88, pp. 289-297, 2018.

[62] S. Zahra and G. George. "Absorptive capacity: A review, reconceptualization, and extension". *Academy of Management Review*, vol. 27, no. 2, Pp. 185-203, 2002.

[63] X. Zhang. "*Understanding Conceptual Framework of Knowledge Management in Government* (*Condensed Version*)". Presentation on UN Capacity-Building Workshop on Back Office Management for e/m-Government in Asia and the Pacific Region, Shanghai, China, 2008.

[64] M. S. Al-Ani, S. N. Jawad and S. Abdelal. "Knowledge management functions applied in Jordanian industrial companies: Study the impact of regulatory overload". *UHD Journal of Science and Technology*, vol. 5, no. 2. pp. 47-56, 2021.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

COVID-19 Classification based on Neutrosophic Set Transfer Learning Approach

Rebin Abdulkareem Hamaamin[1], Shakhawan Hares Wady[1], Ali Wahab Kareem Sangawi[2]

[1]Applied Computer, College of Medicals and Applied Sciences, Charmo University, Chamchamal, Sulaimani, KRG, Iraq, [2]General Science, College of Education and Language, Charmo University, Chamchamal, Sulaimani, KRG, Iraq

ABSTRACT

The COVID-19 virus has a significant impact on individuals around the globe. The early diagnosis of this infectious disease is critical to preventing its global and local spread. In general, scientists have tested numerous ways and methods to detect people and analyze the virus. Interestingly, one of the methods used for COVID-19 diagnosis is X-rays that recognize whether the person is infected or not. Furthermore, the researchers attempted to use deep learning approaches that yielded quicker and more accurate results. This paper used the ResNet-50 module based on the Neutrosophic (NS) domain to diagnose COVID patients over a balanced database collected from a COVID-19 radiography database. The method is a future work of the N. E. M. Khalifa *et al.*'s method for NS set significance on deep transfer learning. True (T), False (F), and Indeterminate (I) membership sets were used to define chest X-ray images in the NS domain. Experimental results confirmed that the proposed approach achieved a 98.05% accuracy rate outperforming the accuracy value acquired from previously conducted studies within the same database.

Index Terms: COVID-19, Chest X-ray, Neutrosophic set, ResNet-50, Classification

1. INTRODUCTION

COVID-19 is a respiratory disease caused by SARS-CoV-2 coronaviruses derive their name from their spherical viruses, which had a shell and a surface projection similar to the solar corona [1]. Unfortunately, the number of deaths from COVID-19 is increasing daily, which has led scientists to work tirelessly to develop a tool to diagnose all types of COVID. There are several ways to diagnose the disease, including blood tests and chest X-ray (CXR) images [2]. The two most popular imaging studies for diagnosing and managing COVID-19 patients are the CXR and computed tomography (CT) scan images. Chest radiography and CT scans, on the other hand, are widely available at most medical centers and typically interpreted with a faster turnaround time than the SARS-CoV-2 laboratory testing. The use of CXR images in the monitoring and examination of numerous lung disorders including tuberculosis, infiltration, atelectasis, pneumonia, and hernia has been known. COVID-19 predominantly affects the respiratory system, resulting in severe pneumonia and acute respiratory distress syndrome in extreme cases. For the most part, X-ray images of the chest are used to diagnose COVID-19-infected patients [3]. Therefore, there are many researches on the diagnosis of COVID-19 using CXR images.

One of the modern methods used in the diagnosis of COVID-19 is the use of deep learning (DL) techniques, which is deep neural network learning. The DL approach has the advantage of automatically extracting features from training data and classifying them more accurately than other traditional methodologies [4]. ResNet, abbreviation for Residual Networks, is a conventional neural network that acts as a foundation for many image processing applications. ResNet's fundamental achievement was that it enabled us to train extraordinarily deep neural networks

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp11-18 **E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Rebin Abdulkareem Hamaamin, Applied Computer, College of Medicals and Applied Sciences, Charmo University, Chamchamal, Sulaimani, KRG, Iraq. E-mail: rebin.abdulkarim@charmouniversity.org

Received: 01-06-2022 Accepted: 17-07-2022 Published: 01-08-2022

with 150+ layers. ResNet-50 architecture is a well-known convolutional neural network (CNN) DL model with 50 layers for image classification [5]. All CXR images are in the spital domain, then transformed into a new domain called the Neutrosophic (NS) domain. The NS includes crisp set, NS graph theory, NS fuzzy set, NS image, and NS topology built on the foundation of NS. The use of these parts on image parts is called advanced image preprocessing, which entails image transformation into the NS domain, The NS domain comprises of three sorts of images, and they are the True (T) images, Indeterminacy (I) images, and Falsity (F) images [6]. All three membership (True, False, and Indeterminate) images were generated in this study.

The identification of COVID-19 as a classification task is addressed in this study using a system based on ResNet-50 architecture in the NS domain. The key contribution of this paper is to analyze the effectiveness of utilizing NS sets based on ResNet-50 architecture using huge database images to improve the overall accuracy and thereby reduce the misclassification error rate. The remainder of the paper is organized as follows. Section 2 presents a review of related research, a complete proposed framework for the detection of COVID-19, including sections such as a database description, image in the NS domain, and ResNet-50 model depicted in Section 3. Section 4 discusses the experimental results and discussions and comparing them with the existing approaches. Finally, Section 5 provides the conclusion of the work.

2. RELATED WORK

Over the past 2 years, numerous studies on COVID-19 infection diagnosis and detection have been conducted. For instance, Lawton [7] proposed COVID-19 detection schema from CT lung scans using transfer learning architectures Standard Histogram Equalization and Contrast Limited Adaptive Histogram Equalization. Five pre-trained CNN-based models (Inception-ResNetV2, InceptionV3, ResNet-50, ResNet101, and ResNet152) for detecting coronavirus pneumonia-infected patients using CXR radiographs were recommended in Narin *et al.* [8] Three separate binary classifications comprising four classes (COVID-19, normal, bacterial pneumonia, and viral pneumonia) were generated applying five-fold cross-validation. Ilyas *et al.* [9] suggested a real-time rule-based Fuzzy Logic Classifier for COVID-19 detection. The suggested methodology collects real-time symptom data from users through an Internet of Things platform to identify symptomatic and asymptomatic COVID-19 patients. Sharmila and Florinabel [10] attempted to classify COVID-19 afflicted individuals using CXR scans utilizing new model of CNN and DCGANs. Hira *et al.* [11] classified COVID-19 patients utilizing nine transfer learning methods. In comparison to other approaches, the ResNet-50 produced the best COVID-19 detection results for binary and multi-classes, according to the outcomes of the experiments.

Saiz and Barandiaran [12] proposed a new testing methodology to determine whether a patient has been infected by the COVID-19 virus using the SDD300 model. The deep feature plus SVM-based procedure was proposed in Singh *et al.* [13] for identifying coronavirus infected patients by applying CXR images. SVM was utilized for classification rather than DL-based classifiers, which require a large database for training and validation. Helwan *et al.* [14] introduced a transfer learning approach to diagnose patients who were positive for COVID-19 and distinguish them from healthy patients using ResNet-18, ResNet-50, and DenseNet-201. For this purpose, 2617 chest CT images of non-COVID-19 and COVID-19 were experimented. Alruwaili *et al.* [15] proposed an improved Inception-ResNetV2 DL model for accurately diagnosing chest CXR images. A Grad-CAM technique was also computed to improve the visibility of infected lung parts in CXR scans. Aradhya *et al.* [16] proposed a system for detecting COVID-19 from CXR scans. In the case of DL architectures, a novel idea of cluster-based one-shot learning was developed. The suggested schema was a multi-class classification system classifying images into four groups: Pneumonia virus, pneumonia bacterial, COVID-19, and typical cases. The proposed schema is built using a combination of an ensemble of Generalized Regression Neural Network and Probabilistic Neural Network classifiers.

Ji *et al.* [17] presented a COVID-19 detection approach based on image modal feature fusion. Small-sample enhancement preprocessing, including spinning, translation, and randomized transformation, was initially conducted using this methodology. Five classic pretraining models including VGG19, ResNet152, Xception, DenseNet201, and InceptionResnetV2 were utilized to extract the features from CXR images. Gaur *et al.* [18] presented an innovative methodology for preprocessing CT images and identifying COVID-19 positive and negative. The suggested approach used the principle of empiric wavelet transformation for preprocessing, with the optimal elements of the image's red, green, and blue channels being learned on the presented approach. Deep and transfer learning procedures recommended by Qaid *et al.* [19] to differentiate COVID-19 cases by assessing CXR images. The designed approaches used either CNN or transfer learning

models to effectively utilize their potential or hybridize them with sophisticated ML procedures Turkoglu [20] presented a pertained CNN-based AlexNet architecture employing the transfer learning technique deployed for COVID-19 identification. The effective features generated using the relief feature selection process were classified using the SVM method at all layers of the architecture. Finally, Al-Ani and Al-Ani [21] reviewed a number of studies on the subject of COVID-19 disease based on a variety of important criteria, including the topic, the applied method, the applied database, the researcher by countries, and the search by country. The findings showed that the majority of research publications supported the claim that coronavirus attacks the human respiratory system.

The rest of this work focuses on expanding and improving the method of applying digital image processing to improve the rate of COVID-19 identification performance using CXR images. The effectiveness of the proposed approach is therefore validated based on different metrics.

3. MATERIAL AND METHODOLOGY

In this paper, an attempt was made to develop a system for the identification and diagnosis of COVID-19 as shown in Fig. 1. To start, all CXR images were cropped to extract only Regions of Interest (ROI) and resized in the preprocessing step. At the second step, the RGB color input images were converted into the NS domains for all three membership subsets. Afterward, the approach divided the NS images through Resnet-50 model into training and testing set in the ratio of 80:20 and then the system runs to use Resnet-50 to classify the CXR images. Finally, the recommended system's performance was evaluated using a variety of well-known metrics including accuracy, sensitivity, specificity, precision, F-score, and Matthews Correlation Coefficient (MCC) rates [22]. The details of the proposed method were presented in the subsequent subsections.

3.1. COVID-19 Database

The structure of the database is the primary stage in any computerized technique. Therefore, a database was created based on the COVID-19 radiography database which is a publicly available database. The database consists of 21165 CXR images, of which 10,192 belong to normal, 3616 correspond to COVID-19 positive, and 6012 belong to lung opacity (non-COVID lung infection) 1345 are labeled as viral pneumonia cases. The data for this paper include 7232 CXR images (Fig. 2), 3616 of which with a positive COVID-19 diagnosis, and 3616 negatives randomly selected to create the balanced database.

3.2. Preprocessing

Image preprocessing refers to the steps performed to prepare images before they are used in model training and validation. Image data preprocessing is the process of converting image

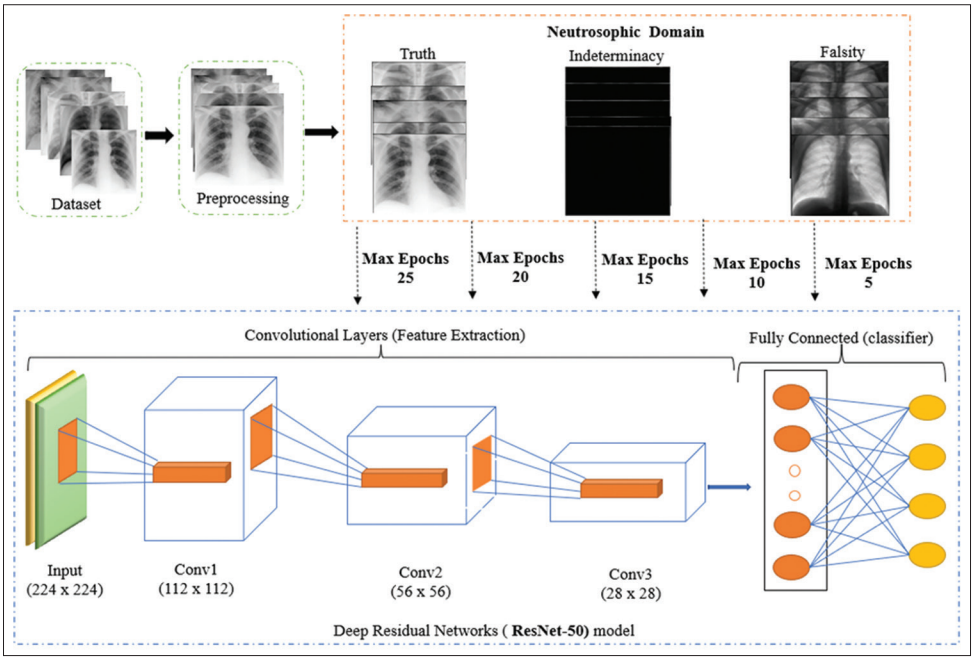


Fig. 1. General architecture of the proposed framework.

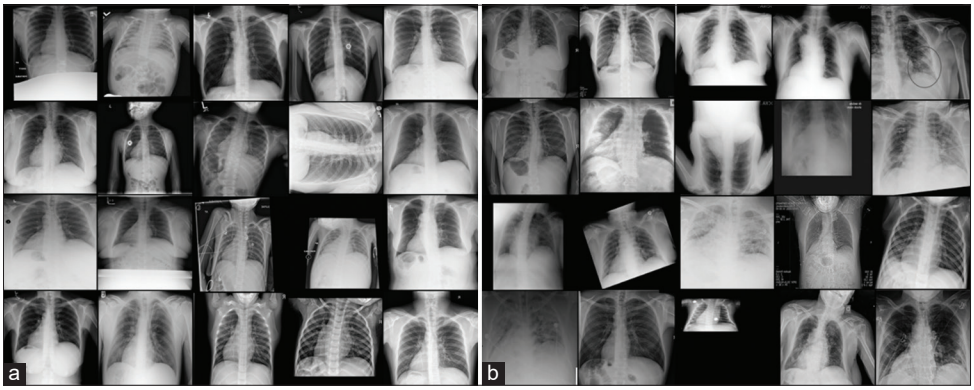


Fig. 2. Examples of CXR images: (a) COVID-19 and (b) Normal.

data into a format that machine learning algorithms can understand. It is widely used to increase the model accuracy while also minimizing its complexities. Image data are preprocessed using a variety of procedures. Therefore, in this step, the bounding box cropping approach is computed to extract the only ROI alone by removing the unwanted background from the input image. Before importing the input CXR images into the proposed framework, the cropped CXR images are resized into fixed size of 256*256 pixels.

3.3. Image in the NS Domain

Neutrosophy is a field of philosophy founded in 1980 by F. Smarandache, which broadened dialectics and investigated the genesis, nature, and extent of neutralities and their interactions with various ideational spectrums. Advanced image processing includes image transformation into the NS domain that includes three areas — background subtraction for foreground objects, edge detection for boundary objects, and background detection for background objects. According to the theory of Neutrosophy, each event has a specific degree of truth (T), falsity (F), and indeterminacy (I), all of which must be taken into account separately. NS truth domain displays all the true parts of the images in percentage. At that point, the image is called an image true. Furthermore, the NS falsity membership degree presents the incorrect parts of the image and become an independent image separate from the other parts. The NS indeterminacy membership degree, which contains the least information of the original image, refers to the uncertain parts of any image [23], [24]. An M * N matrix represents the image as a mathematical object (Spatial Domain). Pixel $P(i,j)$ in the image domain is translated into a NS domain by calculating $PNS(i,j) = T(i,j), I(i,j), F(i,j)$ in equations (1)-(3), where $T(i,j)$, $I(i,j)$, and $F(i,j)$ are taken as probabilities [25] that pixel $P(i,j)$ belongs to white set (object), indeterminate set, and non-white set (background), respectively (Fig. 3).

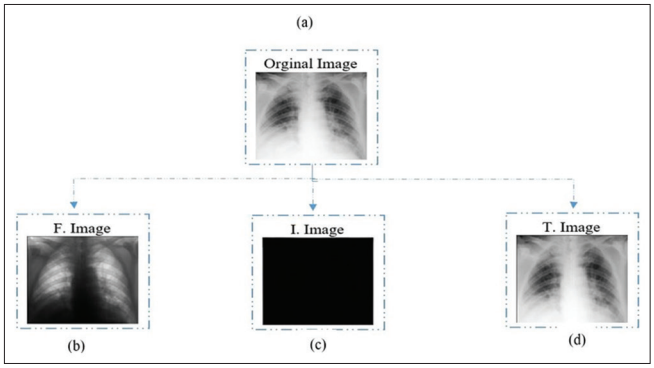


Fig. 3. Convert image from Spatial domain to NS domain: (a) Original Image (COVID-19), (b) F-domain of COVID-19 image, (c) I-domain of COVID-19 image, and (d) T -domain of COVID-19 image.

$$T(i,j) = \frac{\bar{g}(i,j) - \bar{g}_{min}}{\bar{g}_{max} - \bar{g}_{min}} \quad (1)$$

$$I(i,j) = \frac{\delta(i,j) - \delta_{min}}{\delta_{max} - \delta_{min}} \quad (2)$$

$$F(i,j) = 1 - T(i,j) = \frac{\bar{g}_{max} - \bar{g}(i,j)}{\bar{g}_{max} - \bar{g}_{min}} \quad (3)$$

where:

- $g(i,j)$ indicates the pixel intensity value of an image.
- T, I, and F are true, indeterminacy and false sets, respectively, in NS domain.
- $\bar{g}(i,j)$ is the local mean value of $g(i,j)$.
- $\delta(i,j)$ is the homogeneity score of T at (i,j), which is defined as the absolute amount of the difference between an image's intensity value $g(i,j)$ and its local mean value g (i,j).

3.4. Deep Residual Neural Network (ResNet-50) Model

DL is a machine learning method for learning representations that employs artificial neural networks. There are three sorts of machine learning techniques: Supervised, semi-supervised, and unsupervised. Using DL, a computer model learns to do classification tasks directly from images, textual, or numeric data. Deep feature extraction with pre-trained networks such as AlexNet, VGG16, VGG19, GoogleNet, ResNet18, ResNet-50, ResNet-101, InceptionV3, InceptionResNet-V2, DenseNet-201, XceptionNet, MobileNetV2, and ShuffleNet are usually employed for classification tasks [4]. In this research, ResNet-50, which is a better version of CNN, was utilized as the basic model in the architectural proposed design to classify COVID-19 and normal patients' CXR images. The model was pre-trained on the ImageNet database for object detection. ResNet uses shortcuts between layers to reduce interference, which occurs as the network size increases in depth and complexity. With SoftMax activation, the network ends with a 1000 fully connected layer. There are a total of 50 weighted layers with 23,534,592 trainable parameters [26], [27]. The ImageNet database was used to train ResNet-50, a collection of over 14 million images divided into over 20,000 categories designed for image recognition competitions [8].

4. EXPERIMENTAL RESULTS AND DISCUSSION

The critical objective of the proposed framework is to classify CXR images into normal or COVID-19. In this section, ResNet-50 was trained on a dataset containing 7232 images as a benchmark and applied to each subset, namely, CXR images in the NS Domain, by randomly dividing the database into an 80% training set and a 20% testing set. The proposed method was implemented for COVID-19 diagnosis using the MATLAB R2020a programming language on a Windows 10 computer with an Intel Core i7 processor and 16 GB of RAM. The Adam optimizer was used for weight updates, a 1e-4 learning rate, and five epochs; each stage uses the same MiniBatchSize. This method converts images to NS domains with different epochs used for each domain. As Figs. 4-6 depicted the accuracy and loss curves for the three domains.

In addition, experimentations were executed comprehensively to evaluate the performance of the proposed framework in terms of confusion matrix measurements, in particular, the accuracy, sensitivity, specificity, precision, F-score, and MCC rates. A confusion matrix is a table showing how an algorithm classifies data. The structure of the confusion matrix is divided into four parts, Positive True, Positive False, Negative True, and Negative False as shown in Fig. 7. As a result, the images of the database were evaluated in the NS domain with different confidence values for selected models, ResNet-50 model, which was trained with five epochs, Epoch 5, Epoch 10, Epoch 15, Epoch 20, and Epoch 25. Finally, we calculate the average of each domain and compared with the average of other domains.

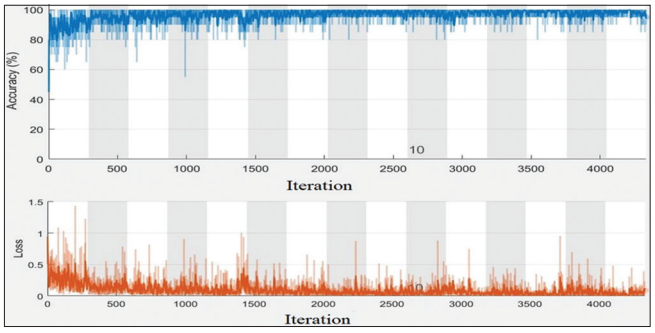


Fig. 4. The accuracy and loss curves of the suggested model that resulted from the T-Domain.

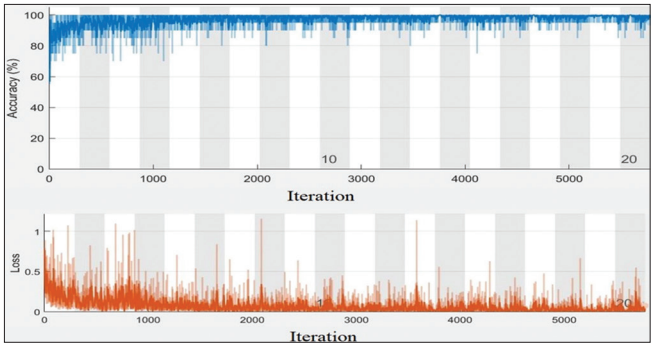


Fig. 5. The accuracy and loss curves of the suggested model resulted from the F-Domain.

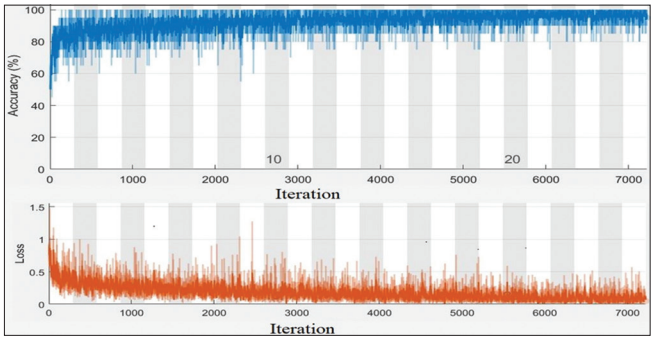


Fig. 6. The accuracy and loss curves of the suggested model that resulted from the I-Domain.

The model achieved an overall accuracy of about 98.05% in the F-domain on the testing set, as shown in Table 1. Furthermore, the highest sensitivity rate acquired in the T-domain had a value of 98.01%, as illustrated in Table 2.

The tables illustrated the model performance evaluation based on specificity, precision, F-score, and MCC, as shown in Tables 3-6. First, the specificity scale of the model yields

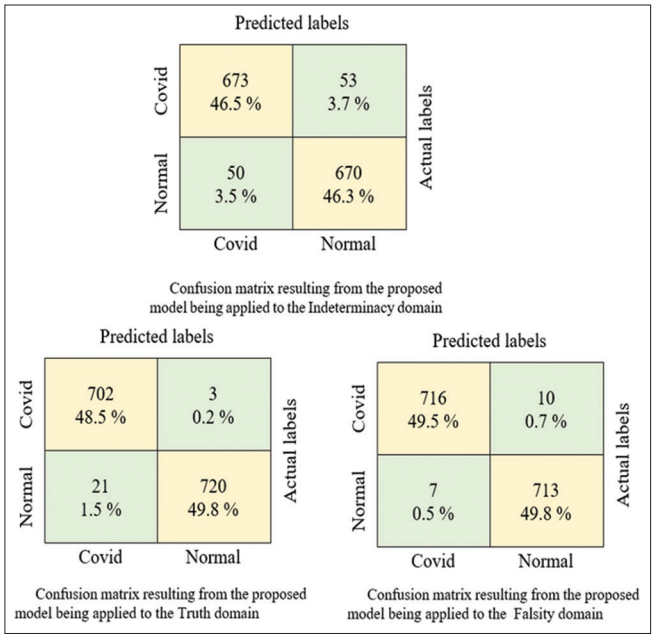


Fig. 7. Confusion matrix of the proposed model.

TABLE 1: Evaluation accuracy of overall epochs

NS Domain	Number of Epoch					
	EP5	EP10	EP15	EP20	EP25	Average (%)
I	87.07	89.21	90.49	89.83	92.32	89.78
T	97.68	97.44	98.34	97.68	97.96	97.82
F	97.51	97.99	98.31	98.55	97.89	**98.05**

EP5: Epoch 5, EP10: Epoch 10, EP15: Epoch 15, EP20: Epoch 20, EP25: Epoch 25, The best result per row is highlighted in bold

TABLE 2: Evaluation sensitivity of overall epochs

NS Domain	Number of Epoch					
	EP5	EP10	EP15	EP20	EP25	Average (%)
I	83.61	88.76	88.81	87.70	91.22	88.02
T	96.83	97.06	99.57	98.05	98.54	**98.01**
F	97.23	96.96	99.23	98.15	96.46	97.61

EP5: Epoch 5, EP10: Epoch 10, EP15: Epoch 15, EP20: Epoch 20, EP25: Epoch 25, The best result per row is highlighted in bold

TABLE 3: Evaluation specificity over all epochs

NS Domain	Number of Epoch					
	EP5	EP10	EP15	EP20	EP25	Average (%)
I	92.71	89.84	92.41	92.23	93.54	92.15
T	98.60	97.84	97.17	97.33	97.43	97.67
F	97.82	99.08	97.42	98.95	99.43	**98.54**

EP5: Epoch 5, EP10: Epoch 10, EP15: Epoch 15, EP20: Epoch 20, EP25: Epoch 25, The best result per row is highlighted in bold

the best result in the F-domain by scoring 98.54% and ultimately outperforming other parts (Table 3). It was found that F-domain improved the highest average prediction specificity and precision to reach 98.54%, whereas the average specificity and precision of the F-domain was the lowest, scoring of 92.15% and 92.42%, respectively (Tables 3 and 4). Furthermore, the same fact has been determined to classify COVID-19 and regular patients CXR images by examining other performance measures (F1-score and MCC) to assess the proposed framework. The outcomes presented that the F-domain reached the maximum F1-score and MCC rates of 98.06% and 96.12% performing ResNet-50, respectively (Tables 5 and 6). From the above experimental results, it is clearly evident that the domain of F has better results in all measures except sensitivity, which effectively discriminates COVID-19 cases from regular patients CXR images more precisely, which may help doctors make a precise diagnosis depending on their clinical specialists as well as the recommended platform as a proper diagnosis tool.

Using the same database and computational environment, the performance of the recommended scenarios was also tested using the misclassification error rate metric. As confirmed in Fig. 8, the misclassification error rates for the recommended scenarios were calculated. The consequences verified that the ResNet-50 Falsity domain results in a minor misclassification error of 1.95% rate, which approved that the proposed scenario outperforms all other scenarios by a significant margin. Therefore, this scenario was considered as a potential classification method for COVID-19 CXR images.

Finally, the proposed system's performance was compared to some state-of-the-art techniques, as shown in Table 7.

TABLE 4: Evaluation precision of overall epochs

NS Domain	Number of Epoch					
	EP5	EP10	EP15	EP20	EP25	Average (%)
I	93.08	89.90	92.74	92.67	93.71	92.42
T	98.62	97.86	97.10	97.30	97.37	97.65
F	97.82	99.10	97.37	98.96	99.45	**98.54**

EP5: Epoch 5, EP10: Epoch 10, EP15: Epoch 15, EP20: Epoch 20, EP25: Epoch 25, The best result per row is highlighted in bold

TABLE 5: Evaluation of F-score over all epochs

NS Domain	Number of Epoch					
	EP5	EP10	EP15	EP20	EP25	Average (%)
I	87.84	89.28	90.71	90.11	92.43	90.07
T	97.71	97.45	98.32	97.67	97.94	97.82
F	97.52	98.02	98.29	98.55	97.93	**98.06**

EP5: Epoch 5, EP10: Epoch 10, EP15: Epoch 15, EP20: Epoch 20, EP25: Epoch 25, The best result per row is highlighted in bold

TABLE 6: Evaluation of MCC over all epochs

NS Domain	Number of Epoch					
	EP5	EP10	EP15	EP20	EP25	Average (%)
I	75.21	78.51	81.10	79.80	84.70	79.86
T	95.40	94.89	96.71	95.37	95.94	95.66
F	95.03	96.01	96.63	97.10	95.84	**96.12**

EP5: Epoch 5, EP10: Epoch 10, EP15: Epoch 15, EP20: Epoch 20, EP25: Epoch 25, The best result per row is highlighted in bold

TABLE 7: Comparison with the current state-of-art/relevant studies

Articles	Techniques	Database	Accuracy %
Afifi *et al*. [28]	CNN-DenseNet161	11,197 CXR images	91.20
Abd Elaziz *et al*. [29]	MobileNetV3+Aqu	21165 CXR images	92.40
Ahmad and Wady [30]	CT, GWT, and LGIP	7232 CXR images	96.18
Walvekar and Shinde [31]	ResNet-50	359 CXR images	96.23
Apostolopoulos and Mpesiana [32]	MobileNetv2	1427 CXR images	96.78
Proposed	ResNet-50+NS	7232 CXR images	**98.05**

The best result per row is highlighted in bold

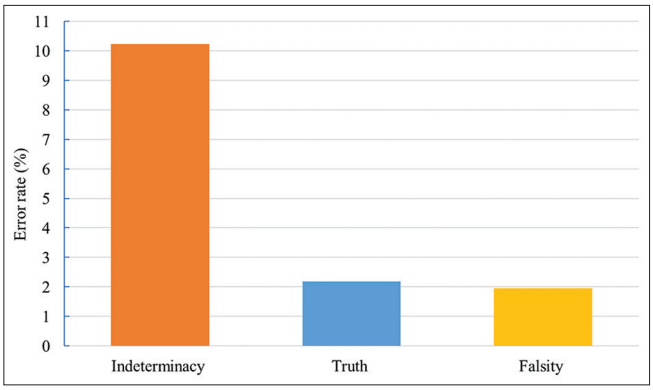


Fig. 8. Misclassification error in the NS Domain.

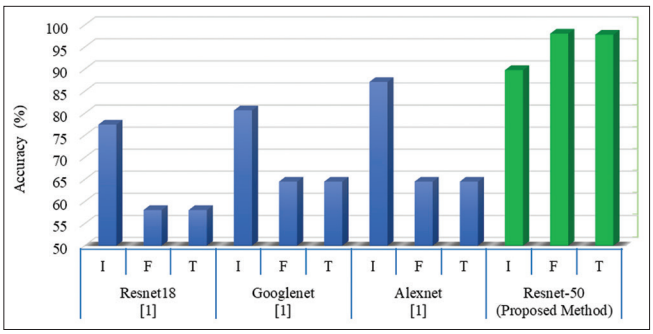


Fig. 9. Comparison of system performance for a different scenario.

Compared to other methods, the proposed system produced excellent outcomes, particularly in average classification accuracy. As a final tool for proposed framework performance evaluation, a comparison has been made with the results obtained from the proposed framework and the results of paper [1] as shown in Fig. 9. The experimentations from Fig. 9 besides obviously confirmed that the proposed system attained the highest result utilizing ResNet-50. This is due to the combined ResNet-50, and NS Domain approaches that helped the model show higher accuracy. Furthermore, the best result was obtained with an overall accuracy of 98.05% compared to the previous studies.

5. CONCLUSION

COVID-19 is the virus that has demolished the world's states and placed everyone under massive quarantine. The virus attacked the world's stability and ushered the world into a new area of instability and chaos. Using technology to control the spread of the virus through the detection of infected patients still requires more work. The work undertaken in this paper aims to serve the people and help the COVID specialist identify the patients more accurately. The study applied the fundamental principles of the NS set. The regulations include True (T) images, Indeterminacy (I) images, and Falsity (F) images on the CXR images database that belonged to both COVID-19 and regular people. Unlike the previous studies, the collected images were transformed into a NS domain trained on the DL technique, and ResNet-50 was used as a transfer learning method to train it on the database. As a result, the model scored 98.05% average accuracy, outperforming other accuracy achieved by the previous studies on a similar database.

REFERENCES

[1] N. E. M. Khalifa, F. Smarandache, G. Manogaran and M. Loey. "A study of the neutrosophic set significance on deep transfer learning models: An experimental case on a limited COVID-19 chest X-ray

dataset". *Cognitive Computation*, vol. 2021, p. 0123456789, 2021.

[2] S. Saadat, D. Rawtani and C. M. Hussain. "Environmental perspective of COVID-19". *Science Total Environment*, vol. 728, p. 138870, 2020.

[3] S. P. Kaur and V. Gupta. "COVID-19 Vaccine: A comprehensive status report". *Virus Research*, vol. 288, p. 198114, 2020.

[4] P. K. Sethy, S. K. Behera, P. K. Ratha and P. Biswas. "Detection of coronavirus disease (COVID-19) based on deep features and support vector machine". *International Journal of Mathematical Engineering and Science*, vol. 5, no. 4, pp. 643-651, 2020.

[5] N. Sharma, V. Jain and A. Mishra. "An analysis of convolutional neural networks for image classification". *Procedia Computer Science*, vol. 132, no. Iccids, pp. 377-384, 2018.

[6] S. F. Ali, H. El Ghawalby and S. A A. "From Image to Neutrosophic Image". In: *Neutrosophic Sets and System*. Port Fuad, Egypt: Port Said University, Faculty of Science, Department of Mathematics and Computer Science Apr. 2015, pp. 1-13.

[7] S. Lawton. "*Detection of COVID-19 from CT Lung Scans Using*". Paper, 2021.

[8] A. Narin, C. Kaya and Z. Pamuk. "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks". *Pattern Analysis Applications*, vol. 24, no. 3, pp. 1207-1220, 2021.

[9] T. Ilyas, D. Mahmood, G. Ahmed and A. Akhunzada. "Symptom analysis using fuzzy logic for detection and monitoring of covid-19 patients". *Energies*, vol. 14, no. 21, p. 7023, 2021.

[10] V. J. Sharmila and J. Florinabel. "Deep learning algorithm for COVID-19 classification using chest X-ray images". *Computational and Mathematical Methods in Medicine*, vol. 2021, p. 9269173, 2021.

[11] S. Hira, A. Bai and S. Hira. "An automatic approach based on CNN architecture to detect Covid-19 disease from chest X-ray images". *Applied Intelligence*, vol. 51, no. 5, pp. 2864-2889, 2021.

[12] F. Saiz and I. Barandiaran. "COVID-19 detection in chest X-ray images using a deep learning approach". *International Journal of Interactive Multimedia And Artificial Intelligence*, vol. 6, no. 2, p. 4, 2020.

[13] A. Singh, A. Kumar, M. Mahmud, M. S. Kaiser and A. Kishore. "COVID-19 infection detection from chest X-ray images using hybrid social group optimization and support vector classifier". *Cognitive Computation*, vol. 2021, p. 0123456789.

[14] A. Helwan, M. K. S. Ma'Aitah, H. Hamdan, D. U. Ozsahin, and O. Tuncyurek. "Radiologists versus deep convolutional neural networks: A comparative study for diagnosing COVID-19". *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 5527271, 2021.

[15] M. Alruwaili, A. Shehab and S. Abd El-Ghany. "COVID-19 Diagnosis using an enhanced inception-resnetV2 deep learning model in CXR images". *Journal of Healthcare Engineering*, vol. 2021, no. 4, pp. 1-16, 2021.

[16] V. N. M. Aradhya, M. Mahmud, D. S. Guru, B. Agarwal and M. S. Kaiser. "One-shot cluster-based approach for the detection of COVID-19 from chest X-ray images". *Cognitive Computation*, vol. 13, no. 4, pp. 873-881, 2021.

[17] D. Ji, Z. Zhang, Y. Zhao and Q. Zhao. "Research on classification of COVID-19 chest X-ray image modal feature fusion based on deep learning". *Journal of Healthcare Engineering*, vol. 2021, pp. 6799202, 2021.

[18] P. Gaur, V. Malaviya, A. Gupta, G. Bhatia, R. B. Pachori and D. Sharma. "COVID-19 disease identification from chest CT images using empirical wavelet transformation and transfer learning". *Biomedical Signal Processing and Control*, vol. 71, p. 103076, 2021.

[19] T. S. Qaid, H. Mazaar, M. Y. H. Al-shamri, M. S. Alqahtani, A. A. Raweh and W. Alakwaa. "Hybrid deep-learning and machine-learning models for predicting COVID-19". *Computational Intelligence and Neuroscience*, vol. 2021, p. 9996737, 2021.

[20] M. Turkoglu. "COVIDetectioNet: COVID-19 diagnosis system based on X-ray images using features selected from pre-learned deep features ensemble", *Applied Intelligence*, vol. 51, no. 3, pp. 1213-1226, 2021.

[21] M. S. Al-Ani and D. M. Al-Ani. "Review study on sciencedirect library based on coronavirus Covid-19". *UHD Journal of Science and Technology*, vol. 4, no. 2, pp. 46-55, 2020.

[22] V. Bahel, S. Pillai. "Detection of COVID-19 Using chest radiographs with intelligent deployment architecture". In: A. E. Hassanien, N. Dey, S. Elghamrawy, editors. *Big Data Analytics and Artificial Intelligence against COVID-19: Innovation Vision and Approach.* Vol. 78. Studies in Big Data, Springer, Cham, 2020.

[23] S. H. Wady, R. Z. Yousif and H. R. Hasan. "A novel intelligent system for brain tumor diagnosis based on a composite neutrosophic-slantlet transform domain for statistical texture feature extraction". *Biomed Research International*, vol. 2020, p. 8125392, 2020.

[24] O. G. El Barbary, R. A. Gdairi. "*Neutrosophic* logic-based document summarization". *Journal of Mathematics*, vol. 2021, pp. 9938693, 2021.

[25] A. Rashno and S. Sadri. "Content-based Image Retrieval with Color and Texture Features in Neutrosophic Domain". In: *3rd International Conference on Pattern Image Analysis IPRIA*, pp. 50-55, 2017.

[26] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos and P. De Geus. "Malicious Software Classification Using Transfer Learning of ResNet-50 Deep Neural Network". In: *Proceeding 16th IEEE International Conference Machine Learning Application*, pp. 1011-1014, 2017.

[27] K. He, X. Zhang, S. Ren and J. Sun. "Deep residual learning for image recognition". 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[28] A. Afifi, N. E. Hafsa, M. A. S. Ali, A. Alhumam and S. Alsalman. "An ensemble of global and local-attention based convolutional neural networks for COVID-19 diagnosis on chest X-ray images". *Symmetry*, vol. 13, no. 1, pp. 1-25, 2021.

[29] M. Abd Elaziz, A. Dahou, N. A. Alsaleh, A. H. Elsheikh, A. I. Saba and M. Ahmadein. "Boosting covid-19 image classification using mobilenetv3 and aquila optimizer algorithm". *Entropy*, vol. 23, no. 11, pp. 1-17, 2021.

[30] F. H. Ahmad and S. H. Wady. "COVID19 infection detection from chest Xray images using feature fusion and machine learning". *The Scientific Journal of Cihan University Sulaimaniya*, vol. 5, no. 2, pp. 10-30, 2021.

[31] S. Walvekar and S. Shinde. "Detection of COVID-19 from CT Images Using resnet50". *In: 2nd International Conference on Communication and Information Processing*, 2020. Available from: https://www.ssrn.com/abstract=3648863 [Last accessed on 2022 Aug 12].

[32] I. D. Apostolopoulos and T. A. Mpesiana. "Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks." *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635-640, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Semantic Web Recommender System over Different Operating Platforms

Halo Khalil Sharif, Kamaran Hama Ali. A. Faraj

Department of Computer Science, College of Science, Sulaimani University, Sulaimani, KRG, Iraq

ABSTRACT

Semantic-Web Recommender System (SWRS) evaluation over different operating systems (OSs) used to facilitate and improve human electronic recommendation management (HERM). The HERM is address the needs of user and dataset of movie in our proposed system through internetworking means which increase the speed of automated recommendation and enhance the goodness of SWRS and services also electronically to select right movies-title to user demand. Furthermore, it will be a benefit for selection a right favor by user for right selection from (i.e., 3000 records in dataset of movie-Lens) in the backend. There are a direct relation between time-consume of selection movie-title, also the time-consume, and accuracy. The two-mentioned parameters, namely, time-consume and accuracy over two different operation system (OSs) which designed by web technology Python. In our research, SWR system is proposed; it is provide with some recommendation methods. The system designed and improved using content-based algorithm (CBA). Investigational results indicate that the developed algorithm technique confident a reasonable performance such as accuracy and time consuming compared to other existing works with a testing average accuracy of 85.63 for windows and 88.35 for Linux operating system. In conclusion, SWRS investigated on two different operating platforms and could be seen that the Linux is faster than windows in accuracy and time consuming.

Index Terms: Semantic Web, E-Recommender System, Content-based, RDF, SPARQL, Python

1. INTRODUCTION

Semantic-Web (as a recommender system) together with all the necessary tools and methods required for creation, maintenance, and application. In actual history, the Semantic-Web is usually future as an heightening of the present World Wide Web (WWW) or (3W) with machine-justifiable data (rather than a large portion of the ongoing Web, which is generally focused on at human utilization), together with services – intelligent agents [1]. Nevertheless, in our proposed system used to facilitate and improve human electronic recommendation management (HERM) is mean that the current web became semantic web (SW) with recommender system (RS). The human consumption artificial intelligent (AI) modify to Semantic-Web Recommender System (SWRS). Our contribution is SW instead of human consumption and RS instead AI and combine to SWRS. However, our proposed system is SW with the cosine similarity is a method and part of content-based algorithm (CBA) for filtering all title-movie in dataset of movie-Lens [2]. The resource description framework (RDF) suggest to graph-based data model, which became part of the Semantic-Web vision [3], the RDF in our proposed system is very necessity with a view to represent data that recommended the title-movie and store into the RDF file. The RDF is much more accurate than the ontology file due to: (1) Easy to use, (2) easy to understand also, nd (3) accurate. Apart from one parameter that used two parameter to enhance accuracy and execution consume-time. The ontology modified from only one parameter to two parameters in propose of

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp21-19-24

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: halo.sharif@univsul.edu.iq

Received: 14-07-2022 Accepted: 28-07-2022 Published: 10-08-2022

the system. Content-based RS make suggestions that consider the users the ratings that users give to items according to their preferences and the content of the items (e.g., extracted keywords, title, pixels, and disk space). The content based algorithms with using the filtering technique is a main idea of our proposed system. The training algorithm is start first for training all dataset to predict the movie-title that situated between limitation and after that, the TEST algorithm is start to filtering of training output. The activates are depend on training algorithm and TEST algorithms between the user's demands and movie's title (plus demands) to build the SWRS decisions. Semantic-Web utilizes the Resource Description Framework (RDF) and the Simple-Protocol and Query/Update Languages (SPARQL) as uniform logical data illustration and handling models, permitting machines to straight interpret data from the Web. As Semantic-Web, applications is growing progressively popular, new-fangled and stimulating threats of security arise [4], it is impossible to achieve our proposed system or any evaluation without RDF because in RDF is store and transfer data to web application through SPARQL. The two parameters that mentioned namely tag line and original title. Nevertheless, the only used parameter is overview parameter used in Cami *et al.* [5]. The contribution in our proposed system is two parameters. While the deployment of (www) and the internet was swiftly increasing, the recommendation outfits become electronic to support e-commerce (EC) business. Usually, the concept of E-recommender is relevant with all kinds of digitalizes businesses and it uses three-tier architecture [6]. Regardless of the fantastic measure of data that is accessible in the reality or on the Web, it is difficult for the searcher to track down items or services that he may be interested in. Decision-making is an essential part that the traditional and electronic recommendation should do. The vast amounts of digitally available candidate information denote a sizeable opportunity for improving matching quality and it leads to better web semantic recommendation performance [7]. This paper proposes a new procedure for recommending movie-titles using a content-based filtering algorithm and generally used dataset (MovieLens). The whole of the paper is arranged as follows. Section 2 places forward a literature review. Section 3 shows a complete SWRS for the recommending of movie-titles, containing units like an outline of system architecture, MovieLens dataset description, data preprocessing, feature extraction, and performance metrics. Section 4 discusses the experimental results achieved after applying different feature extractors and comparing them from different platforms with the existing methods. Finally, Section 5 deals with the conclusion of the work.

2. LITERATURE SURVEY

In Semantic-Web Recommender System (SWRS), techniques have conveyed exceptional outcomes; these techniques are regularly acted in the recommendation on movie-titles dataset. Recently, various works were executed with the assistance of various content-based methodology to distinguish and predict of movie-titles. A short audit of a few significant contributions from the current literature is given.

Soumya Prakash Rana (2020) [8] proposed arrangement, health recommender systems (HRS) have arisen for patient-situated decision-making to suggest better medical care guidance in light of profile health records (PHR) and patient data sets. The HRS can upgrade medical services frameworks and at the same time oversee patients experiencing a scope of various sicknesses utilizing prescient investigation and suggesting fitting therapies. A content-based recommender system (CBRS) is a tweaked HRS approach that focuses on the assessment of a patient's set of experiences and "learns," through AI (ML), to produce forecasts. Moreover, CBRS plans to offer personalized and believed data to the patient's with respect to their health status.

Donghui Wang (2018) [9] they fostered a content-based diary and meeting recommender framework for software engineering and innovation. To the extent that, there is no comparative recommender system or distributed strategy like what they have presented here. Besides, there was no dataset to utilize. Hence, the web crawler has been intended to gathering information and creates preparing and testing informational indexes. Then, unique component determination techniques and played out few trials used to choose a decent system and recreate include space. Despite the fact that accomplishing 61.37% exactness for paper proposal.

Ibukun Tolulope Afolabi (2019) [10], in this examination, showed a semantic-web content digging approach for recommender frameworks in web based shopping. The strategy depends on two significant stages. The primary stage is the semantic preoperational of text-based information utilizing the blend of a created cosmology and a current metaphysics. The subsequent stage utilizes the Naïve Bayes calculation to make the proposals. The result of the framework is assessed utilizing accuracy, review and f-measure.

Carlos Luis Sanchez Bocanegra (2017) [11] this shows the practicality of utilizing a semantic content-based recommender framework to enhance YouTube health

recordings. Assessment with end-clients, notwithstanding medical services experts, will be expected to distinguish the acknowledgment of these suggestions in a no simulated data looking for setting. Most of sites suggested by this framework for health recordings were pertinent, in view of evaluations by health experts.

Albatayneh (2018) [12], this examined to present an original proposal engineering that can prescribe intriguing post messages to the students in an e-learning on the web conversation gathering in view of a semantic content-based separating and students' negative appraisals. We assessed the planned e-learning recommender framework against leaving e-learning recommender frameworks that utilization comparable sifting methods concerning suggestion exactness and students' exhibition. The got exploratory outcomes display that the suggested e-learning recommender framework beats other comparative e-learning recommender frameworks that utilization non-semantic content-based separating strategy (CB), non-semantic content-based sifting method with students' negative appraisals (CB-NR), semantic content-based sifting procedure (SCB), concerning framework precision of around 57%, 28%, and 25%, separately.

3. PROPOSED METHODOLOGY

3.1. System Architecture

TF-IDF is used for the vectorization of the information and cosine similarity is utilized to compute the similarity measure between the vectors. TF-IDF is normally used as a portion of content-based algorithm recommendations systems in proposed system. It contains of two positions: Term-Frequency (TF) and Inverse-Document-Frequency (IDF). TF deals-with the occurrence of interests and preferences in user profile. Whereas, IDF deals with inverse of the word frequency among the entire data provided by user profile. These two theories are joint together to present the recommendation for a user based on the data's presented by user profile. Cosine similarity be able to catch the similarity among two attribute or more from the dataset found by determining cosine value between two vectors or more. Use of cosine similarity can be executed on any two texts such as documents, sentences, attributes or paragraph. Occasionally through the similarity measurement between the vectors which produce unstable results. Finally, the SWRS are build using famous algorithm content-based (CB) and RDF. The important steps in proposed structure design are shown in Fig. 1. In the below figure shoe all steps as an instruction of our system. ROW one show all main steps, but the underneath RAW is subset of first RAW. RAW one and two are complete each other's for the sake of processes of the system.

3.2. Dataset Explanation

The proposed system was trained as well as tested on the MovieLens dataset. The dataset consists of movies released on or before July 2019. Information focuses incorporate cast, group, plot, watchwords, spending plan, income, banners, delivery dates, dialects, creation organizations, nations, TMDB vote counts, and vote midpoints. The Complete MovieLens Datasets comprising 26 million evaluations and 750,000 label applications from 270,000 clients on every one of the 45,000 motion pictures. This dataset is a troupe of information gathered from TMDB and GroupLens. The Movie Detail, Credit and Keyword have been gathered from the (TMDB) open an API. This item utilizes the TMDb API however is not embraced or affirmed by TMDb. Their API likewise gives admittance to information on numerous extra motion pictures, entertainers and entertainers, group individuals, and TV shows. The Movie Links and Ratings have been gotten from the Official GroupLens site. A portion of the things you can do with this dataset: Predicting film income or potentially film achievement in view of a specific measurement. What motion pictures will generally get higher vote counts and vote midpoints on TMDB? Building Content-Based and Collaborative Filtering Based Recommendation Engines [13].

3.3. Preprocessing

To be capable handling information concurring appropriately, really, and productively, that it requires the capacity as far as

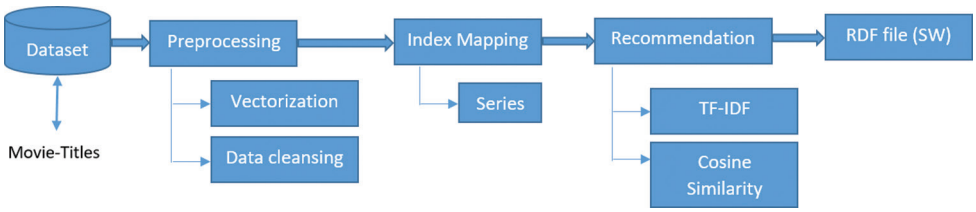


Fig. 1. Architecture of proposed system.

a specific programming language that is explicitly devoted to handling information or data in many place of origin in the association or the web to turn into a valuable information researcher for associations or organizations [14], because of in the proposed technique (fillna) method is used to cleansing data from the dataset to achieve the best result. Scikit-learn is a permitted software (utility) machine-learning library for the Python programming language. It assists python numerical and scientific libraries, in which Tfidf-Vectorizer is one of them. It alters a group of raw documents to a matrix of TF-IDF structures. As tf–idf is extremely frequently used for text sorts, the class Tfidf-Vectorizer merges all the options of Count-Vectorizer and Tfidf-Transformer in a particular model. Tfidf-Vectorizer uses an in-memory vocabulary (a python dict) to map the most recurrent words to features indices and henceforward calculate a word occurrence frequency (sparse) matrix, the class of TfidfVectorizer used to vectorizing the two attribute from the dataset (MovieLens) in our proposed system.

3.4. Recommendation Engine

Term-Frequency Inverse-Document-Frequency (TF-IDF) is utilized to yield recommendations to the user's favorites. Each data attribute from the datasets is converted into a vector by applying the TF-IDF vectorization algorithm described before. For each vector, a similarity measure is calculated using the cosine similarity method. When a user requires number of recommendations for a certain movie, the correspondence quantities are produced for the movies with concern to that movie. Individually similar movie detected will have a confident score of how similar it is to the represented movie, which is sorted into descending order, because of list the movies with high to low similarity. Conferring to the amount of recommendations demanded by the user, the indices of those movies are gathered and showed to the user as a list of movies. The recommendations created by the engine are displayed over a user interface to the user; the engine is trained to yield similarity measures using the training data. The backend is scripted using Python language, whereas the calculations performed from Equations 1 to 4 to find Cosine-Similarity and TF-IDF [15].

Cosine similarity, Based on vector similarity, similarity among vectors can be denoted as Eq. 1:

$$cos(\theta) = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} Ai^2}\sqrt{\sum_{i=1}^{n} Bi^2}} \quad (1)$$

Where, Ai and Bi are components of vector A and B respectively:

TF, i.e. word frequency, indicates the frequency of terms in the text showed in Eq. 2.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k} n_{i,j}} \quad (2)$$

Where, n is the frequency of terms in the movie-title.

IDF, i.e. inverse document frequency, represents the reciprocal of the quantity of movie-title containing words in the mass displayed in Eq. 3.

$$idf_{j=log\left[\frac{n}{df_j}\right]} \quad (3)$$

Where, n is the frequency of movie-title containing words

Thus, the TF-IDF weight for catchphrase in record can be composed in Eq. 4

TF-IDF = (Frequency of words/Total words of sentences) × (Total documents/Documents containing the word) (4)

3.5. Evaluation

Evaluation is used to assessment the consideration space and results from various models or algorithms. For the recommendation of movie-titles, so when it comes to a classification problem, can be counted on an AUC - ROC Curve. Because of needed to scan or imagine the performance of the proposed system, It is denoted by the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. It is one of the greatest significant estimation metrics for testing any arrangement model's performance. It is as well written as AUROC (Area Under the Receiver Operating Characteristics).The range AUC is between 0 and 1, An brilliant model has AUC proximate to the 1 and that implies it has a moral degree of distinguishability. The unwell model has an AUC near 0 which denoted it has the poorest measure of separability. Three broadly utilized performance metrics were applied to assess the proposed system's performance: TPR (True Positive Rate)/Recall/Sensitivity, Specificity and FPR(False Positive Rate)/Precision. To calculate the metrics specified by Equations 5–7, three distinct performance factors were selected: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

TPR (True Positive Rate)/Recall/Sensitivity:

$$TPR / Recall / Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

Specificity:

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

FPR:

$$FPR = \frac{FP}{TN + FP} \quad (7)$$

4. RESULTS AND DISCUSSION

Although the Semantic-Web Recommendation System (SWRS) built by other developers have used any technique filtering techniques, they had encountered weaknesses, which were slight disturbing. In our paper, we had implemented (SWRS) by content-based algorithm in two attributes from the MovieLens dataset utilizing Cosine-Similarity and Term-Frequency Inverse Document-Frequency (TF-IDF), after that the algorithm is tested on the windows 10 64-bit and Linux 18.2 64-bit operating system with the different number of records (movie-title), then these results are shown in Table 1 that display the results achieved on windows 10 64-bit operating system in different number of records in our dataset to produce process time, execution time and accuracy form read dataset to create RDF file, furthermore Table 2 that display information as Table 1 but on the real (not virtual) Linux 18.2 64-bit operating system. These marks pointed to that the building of (SWRS) on Linux operating system is better than on windows operating system, moreover the Fig. 2 on windows 10 and Fig. 3 on Linux operating system demonstrates to verify the results that found by Area Under Curve (AUC) to accuracy of creating SWRS. As a result of all the evaluation found out the two parameters are better in Quality of service (QoS), Quality of information (QoI), in spite of the results (accuracy and speed) can be affected by the features of the computer such as (CPU, RAM, Data Bus, Graphic Card) for this situations, so these issues should be handled before any processing to provide the predicted results.

TABLE 1: Results of the SWRS on windows-V10 operating system

Windows 10 64-bit Operating System			
No. Records in Dataset	Process Time (Second)	Execution Time (Second)	Accuracy (AUC)
1000	1.00315	1.01364	88.75%
2000	1.07825	1.15712	88.25%
3000	1.40268	1.52974	87.15%

TABLE 2: Results for the SWRS on linux-V22 operating system

Linux V 18.2 64-bit Operating System			
No. Records	Process Time (Second)	Execution Time (Second)	Accuracy (AUC)
1000	0.7104	0.6034	92.10%
2000	0.7445	0.6499	91.75%
3000	0.8268	0.6726	90.35%

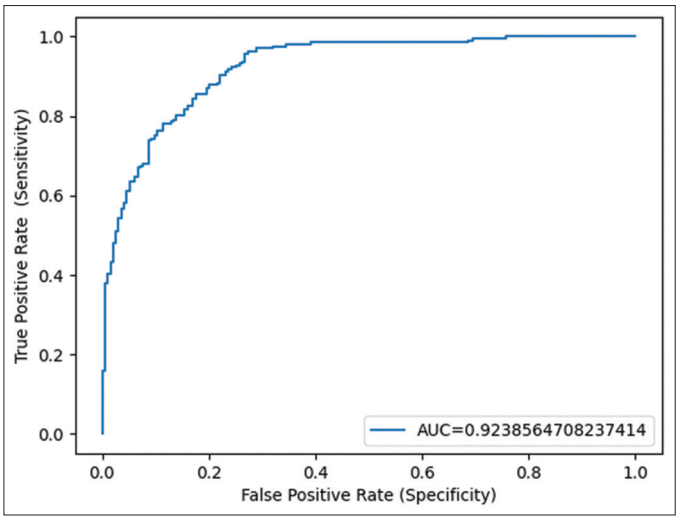


Fig. 2. Display AUC on windows for SWRS.

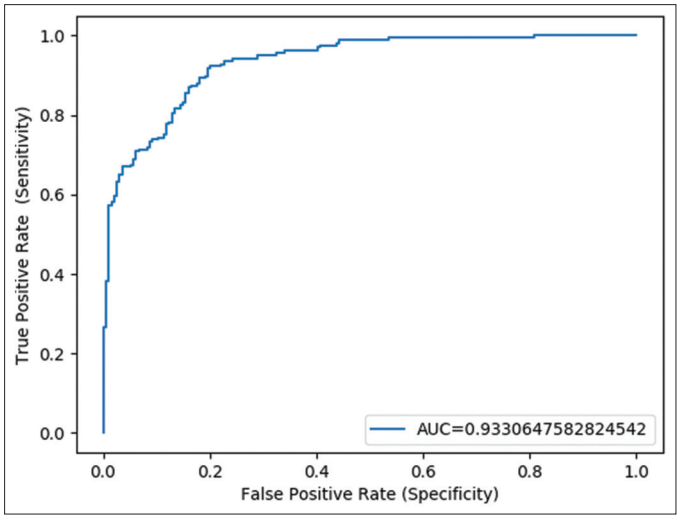


Fig. 3. Display AUC on Linux for SWRS.

5. CONCLUSIONS

Quick recommendations of movie-title, provides the greatest fortuitous to finding the correct titles (movies) to the users, semantic web-based content-based algorithm recommender system able to use in automatically and successfully analyze required data to identify the movie-titles. The main objective of our research is to use TF-IDF and cosine Similarity model to perform recommendations then creating RFD file as semantic web using input data from the amount of output of the data that recommended by the proposed method, after that Simple Protocol and RDF Query Language (SPARQL) used and it is the query language for the Semantic-Web that performs demand information from the databases or any data source that can be plotted to RDF. The proposed system offered a higher average recommendation accuracy approximately (88.5%) for windows operating system and (91.25%) for Linux operating system investigational results exposed that the proposed method is more effective than the previous works.

REFERENCES

[1] P. Hitzler, 2021. "A review of the semantic web field". *Communications of the ACM*, vol. 64, pp.76-83, 2021.

[2] I. Portugal, P. Alencar and D. Cowan. "The use of machine learning algorithms in recommender systems: A systematic review". *Expert Systems with Applications*, vol. 97, pp. 205-227, 2018.

[3] J. E. Gayo, E. Prud'hommeaux, I. Boneva and D. Kontokostas. "Validating RDF data". *Synthesis Lectures on Semantic Web*: *Theory and Technology*, vol. 7, pp. 1-328, 2017.

[4] H. Asghar, Z. Anwar and K. Latif. "A deliberately insecure RDF-based semantic web application framework for teaching SPARQL/ SPARUL injection attacks and defense mechanisms. *Computers and Security*, vol. 58, pp. 63-82, 2015.

[5] B. R. Cami, H. Hassanpour and H. A. Mashayekhi. "A content-based Movie Recommender System Based on Temporal User Preferences". *In: 3rd Iranian Conference on Intelligent Systems and Signal Processing* (*ICSPIS*). pp. 121-125, 2017.

[6] G. M. Zebari, K. Faraj and S. Zeebaree. "Hand writing code-php or wire shark ready application over tier architecture with windows servers operating systems or linux server operating systems". *International Journal of Computer Sciences and Engineering*, vol. 4, pp. 142-149, 2016.

[7] K. Faraj. "*Design of an E-commerce System Based on Intelligent Techniques*". Sulaimani University, Sulaimani, KRG, Iraq, 2010.

[8] S. P. Rana, M. Dey, J. Prieto and S. Dudley. "Content-based Health Recommender Systems". In: *Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries*. John Wiley and Sons, Hoboken, pp. 215-236, 2020.

[9] D. Wang, Y. Liang, D. Xu, X. Feng and R. Guan. "A content-based recommender System for computer science publications". *Knowledge-Based Systems*, vol. 157, pp. 1-9, 2018.

[10] I. T. Afolabi, O. S. Makinde and O. O. Oladipupo."Semantic web mining for content-based online shopping recommender systems". *International Journal of Intelligent Information Technologies*, vol. 15, pp. 41-56, 2019.

[11] C. L. Bocanegra, J. L. Ramos, A. Civitet and L. F. Luqure. "HealthRecSys: A semantic content-based recommender system to complement health videos". *BMC Medical Informatics and Decision Making*, vol. 17, pp. 1-10, 2017.

[12] N. A. Albatayneh, K. I. Ghauth and F. F. Chua. "Utilizing learners' negative ratings in semantic content-based recommender system for e-learning forum". *Journal of Educational Technology and Society*, vol. 21, pp. 112-125, 2018.

[13] Available from: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset [Last accessed on 2022 Feb 05].

[14] S. Sardjono, R. Y. Alamsyah, M. Marwondo and E. Setiana. "Data cleansing strategies on data sets become data science". *International Journal of Quantitative Research and Modeling*, vol. 1, pp. 145-156, 2020.

[15] R. H. Singh, S. Maurya, T. Tripathi, T. Narula and G. Srivastav. "Movie recommendation system using cosine similarity and KNN". *International Journal of Engineering and Advanced Technology*, vol. 9, pp. 556-559, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Newly Simple Quantitative Determination of Montelukast Sodium by Ultraviolet-Spectrophotometry

Dlivan Fattah Aziz[1], Yehia Ismail Khalil[2]

[1]Department of Pharmaceutics, College of Pharmacy, University of Sulaimani, Sulaimanyia, Iraq, [2]Department of Pharmaceutics, College of Pharmacy, University of Baghdad, Baghdad , Iraq

ABSTRACT

Montelukast sodium is well known pharmaceutically for its action as leukotriene antagonist and reliving symptoms associated with asthma is available in the market as tablet, chewable tablet, and powder. The aim of this study was to develop newly simple selective ultraviolet spectrophotometry (UV) method for daily routine analysis of quality control department. The UV method was developed with wavelength at 287.0 nm. This newly developed method was effectively applied to tablet dosage form of the motelukast sodium follow the Beer's Lamberts at range 2.5–50 µg/mL. The validated parameters were carryout such as linearity, accuracy, precision, and specificity. The result of validation statistically studied and found to be satisfactory.

Index Terms: Ultraviolet, Montelukast, Determination, Validation, Quantitative, Method

1. INTRODUCTION

Montelukast sodium (MTK) which has the following chemical structure (Fig. 1) is considered as a good alternative to corticosteroid inhaler in treating asthma and rhinitis since it has fewer side effects [1]. MTK mechanism of action is by blocking the action of CysLT receptor Type 1 in respiratory system that results in relaxing smooth muscle and decreasing inflammation. MKT hydrophobic acidic drug that has water solubility about 0.2–0.5 µg/mL at room temperature; therefore, it is considered as Class II compound according to biopharmaceutic category system [2].

Montelukast base solubility enhanced through salt formation as sodium salt of montelukast MTK. MTK possess acidic lipophilic property with a PKa between 2.7 and 5.8 and logP 8.79 which make it soluble in higher pH media [2]. MTK is available as a tablet dosage form under the brand name of Singulair for both adult and children from age 6 months and older with no detected adverse effect [3].

Different methods have been studied to determine amount of MTK in its dosage form such as capillary electrophoresis [4], cyclic voltammetry [5], high performance liquid chromatography (HPLC) with florescence detection [6], and HPLC with ultraviolet (UV) detection [7], this study develop simple, specific, accurate, and precise method by UV-spectrophotometry and validate it according to International Council for Harmonisation (ICH) guideline, and evaluate this new method with previously published method that has the same way of determination.

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp24-28

E-ISSN: 2521-4217

P-ISSN: 2521-4209

2. MATERIALS AND METHODS

2.1. Instrumentation

For determination UV double beam (Spekol 2000, analytikjena, Canada) with two identical 1 cm quartzes cell

Corresponding author's e-mail: Dlivan Fattah Aziz, Department of Pharmaceutics, College of Pharmacy, University of Sulaimani, Sulaimanyia, Iraq. E-mail : dlivan.aziz@univsul.edu.iq

Received: 03-04-2022 Accepted: 01-08-2022 Published: 15-08-2022

was used, all materials were weighed by electronic sensitive balance (Sartorius, Germany), water bath sonicator (Starsonic, Italy) used to aid dissolving solute in solvent during solution preparation.

2.2. Materials

Pure MTK, lactose monohydrate, magnesium stearate, microcrystalline cellulose, and croscarmelose sodium were kindly provided by PiONEER for pharmaceutical industry, Iraq. Ethanol 96% was purchased from Merck, Germany.

2.3. Standard Stock Solution and Calibration Curve Solution Preparation

Stock solution of 100 μg mL^{-1} of active pharmaceutical ingredient (API) was prepared by dissolving 25 mg of API into 250 mL of diluent (Water: Ethanol) (1:1 V/V) sonicated for few minutes. Series solutions of the following concentration were prepared from stock solution in same the diluent (2.5, 5, 10, 15, 25, and 50 μg/mL). Each solution was read triplicate and average of each sample was put into linear graph.

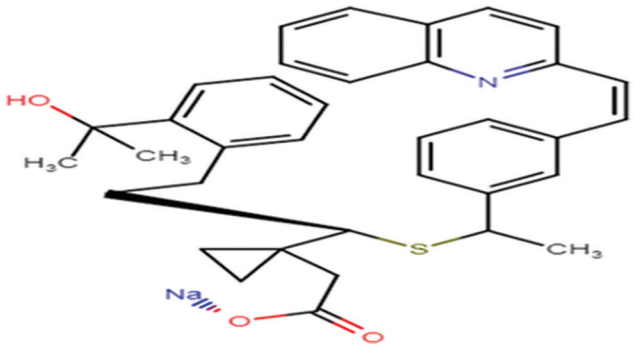


Fig. 1. Chemical structure of montelukast sodium.

2.4. Method Development

Different media for dissolving MTK were evaluated to choose best solvent for API depending on solubility of MTK, stability, cost, selectivity, and toxicity. First water used as solvent then ethanol was added gradually till found that ethanol with water by 1:1 (V/V) will give clear solution. The prepared standard solution scanned and found that best absorption would be at 287.0 nm.

2.5. Stability

Stability of MTK solution of calibration was determined at room temperature in day light condition for a period of 24 h by observing change in absorbance at the same wavelength.

2.6. Analytical Validation

ICH (Q2) R guideline of validation of analytical procedure was applied for validating developed method as the following:

2.6.1. Precision

Both interday and intraday of precision were analyzed with median concentration of API. Intraday precision was completed by evaluating the median concentration of the MTK at the

TABLE 1: Different concentration of MTK solution at different level with corresponding absorption

Linearity sample 2 zero time					
level	**Conce ug/mL**	**abs**	**avearge Area**		
50%	2.50	0.1136	0.1138666667	slope	0.04699946417
		0.114		intercept	0.01064293369
		0.114		r	0.9996766948
				r2	0.9994
75%	5.00	0.2447	0.2449		
		0.245			
		0.245			
100%	10.00	0.4784	0.4784		
		0.4784			
		0.4784			
125%	15.00	0.7093	0.7093		
		0.7093			
		0.7093			
150%	25.00	1.225	1.226533333		
		1.2273			
		1.2273			
200%	50.00	2.3433	2.3433		
		2.3433			
		2.3433			

same day while interday precision was studied over consecutive days for the same concentration repeated 6 times. Evaluating precision of an analytical procedure provide statistics data on the unsystematic error. It states agreement between numbers of measurements achieved from several sampling of the same identical sample under approved conditions. The percentage of relative standard deviation (% RSD) values were studied and the low value of % RSD indicated the precisely of the analytical procedure. The value % RSD for the precision study according to ICH guideline should be <2% (interday precision) this to confirms good precision of the method.

2.6.2. Recovery

Recovery was completed using the method where identified quantity of standard MTK equivalent to 75, 100, and 125% of linear concentration had been added to placebo. The samples were read 3 times and percentage amount of API was calculated at each level.

2.6.3. Linearity

Calibration curve for standard MTK solution was obtained in range from 2.5 μg mL^{-1} to 50 μg mL^{-1} for MTK. Peak absorbance for each concentration must plot against respective concentrations and linear regression analysis should obtain the correlation coefficient higher than 0.999 to confirm that there is an excellent relationship between the absorbance and concentration of the samples and method have linear in response.

2.7. Statistical Analysis

Basic statistical analysis was applied such as mean, standard deviation, average, and RSD% using Microsoft Excel.

3. RESULTS AND DISCUSSION

This rapid technique for determination of the MTK is useful in drug analysis especially in pharmaceutical industry when time costs especially using HPLC for determination of MTK is time consuming and requires effort and high cost. Choosing the best solvent for preparing solution of MTK is a bit challenging in this study. Different solvents have been tried and found that equal volume of (water: ethanol) give clear solution, its cheap, available in almost every laboratory, and easy to use. This diluent makes this study different from other previously studies in which in most of them, Methanol 100% [8]-[10], Methanol 50% [11], [12], methanol with 0.1N NaOH [13], chloroform [14], or 7.4 pH phosphate buffer with 0.5% sodium lauryl sulfate [15] were used during solution preparation of MTK.

TABLE 2: Precision of the method (*n*=6).

Parameter	Amount obtained by the proposed method in (mg)	
	Interday	Intraday
Mean	10.13	10.44
SD	0.00923	0.0105
RSD%	0.91%	1.01%

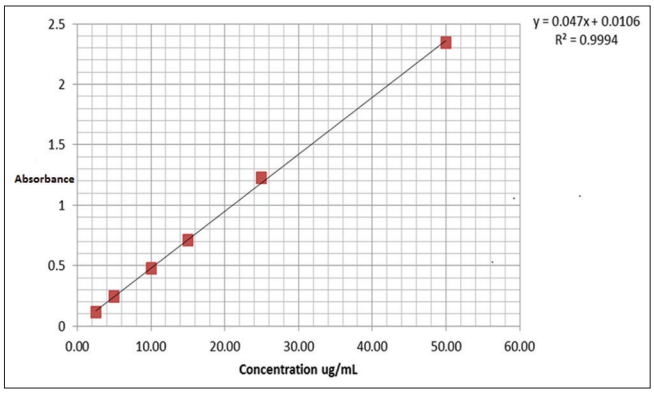


Fig. 2. Calibration curve of MTK at different concentration.

Different wavelengths have been suggested for reading MTK standard solution in different articles such as 344.4 [11], 359 [8], 344.3 [10], 283 [9], 287.3 [15], 286.5 [13], and 280 nm [12] but in the present research when standard solution of MTK in equal mixed volume of (ethanol: water) scanned by UV absorbance to find characteristic peak at wavelength between 400 and 200 nm by 0.2 nm interval, it was found that best absorbance is at 287.0 nm at nominal concentration of 10–15 μg/mL. The suggested method was validated regarding to ICH guideline in the following aspects.

3.1. Specificity and Selectivity

MTK solution of concentration 10 μg/mL in diluent was prepared in both alone and mixed with common excipient such as (lactose monohydrate, magnesium stearate, microcrystalline cellulose, and croscarmelose sodium) separately to know the interference of these excipients with API. Both solutions were scanned at wavelength between 400 and 200 nm. Method was specific and selective and there was no interference in reading between API and excipients.

3.2. Linearity

For linearity, six different concentrations were prepared from lower concentration 2.5 μg/mL to higher concentration 50.0 μg/mL. Each concentration was read 3 times as shown in the Table 1. Linear relationship was observed between

TABLE 3: Accuracy of MTK

Standard Data						
Weight	**dil. Factor**	**Conc. μg/mL**	**Area**	**Average Area**	**SD**	**RSD%**
		10.000	0.47840 0.47840 0.47840	0.4784	0.00	0.0%
Accuracy - recovery of MTK						
Level	**Quantity spiked μg/mL**	**Area**	**Quantity recovered μg/mL**	**% recovery**	**Average Recovery (%)**	**%RSD**
75	5.00	0.2395	5.01	100.13	100.25	0.11
		0.2400	5.02	100.33		
		0.2399	5.01	100.29		
100	10.00	0.4760	9.95	99.50	99.45	0.04
		0.4757	9.94	99.44		
		0.4756	9.94	99.41		
125	15.00	0.7105	14.85	99.01	99.04	0.04
		0.7110	14.86	99.08		
		0.7107	14.86	99.04		
Average					99.58	

concentrations of MTK besides mean reading of absorbance at each point as it is clear in Fig. 2 the determination correlation coefficient (γ^2) equal to 0.999.

3.3. Precision

The method was evaluated to confirm precise in repeatability of analyzing six samples. The samples were prepared and the percentage of label claim of API of each sample was statistically evaluated. Results are shown in Table 2. The results were accepted according to acceptance criteria for assay value obtained from single analyst %RSD should be <2.0% while %RSD of two analyst performing the same samples that terminated in both days should not be more than 3.0%.

3.4. Recovery

Accuracy and recovery of assay for the method was demonstrated by analyzing data achieved from standard addition into placebo solution at three levels. The amount of recovery of each sample was determined in percentage at each level and % RSD was calculated that was <2.0% shows a good accuracy of method. As it is clear in Table 3, according to ICH guidelines, good recovery of API should lie within the range of 98–102% which means the percentage recovery of API added to placebo should be in range of 100 ± 2.0% for average of three weight samples at each level.

3.5. Stability

The prepared solutions were stored at room temperature 25°C, analyzed after 24 h and it was found that MTK is stable for this period of analysis in diluent.

4. CONCLUSION

The validated UV method for MTK determination indicated that the method is linear, accurate, rapid, and specific. The simplicity of method allows it to be used in laboratories that have simple equipment and lack HPLC, liquid chromatography mass spectrometry, or ultra-performance liquid chromatography especially for repetitive analysis of MTK in pharmaceutical dosage form or during development of new dosage form of MTK. The current method is also useful for quantitative determination of MTK in quality control department in pharmaceutical industry.

5. ACKNOWLEDGMENT

The author would like to thank PiONEER pharmaceutical industry for providing facility.

REFERENCES

[1] N. Kittana, S. Hattab, A. Ziyadeh-Isleem, N. Jaradat and A. N. Zaid. "Montelukast, current indications and prospective future applications". *Expert Review of Respiratory Medicine*, vol. 10, pp. 943-956, 2016.

[2] S. Sawatdee, T. Nakpheng, B. T. W. Yi, B. T. Y. Shen, S. Nallamolu and T. Srichana. "Formulation development and *in-vitro* evaluation of montelukast sodium pressurized metered dose inhaler". *Journal of Drug Delivery Science and Technology*, vol. 56, pp.101534, 2020.

[3] C. Cingi, N. B. Muluk, K. Ipci and E. Şahin. "Antileukotrienes in upper airway inflammatory diseases". *Current allergy and asthma reports*, vol. 15, pp. 1-11, 2015.

[4] Y. Shakalisava and F. Regan. “Determination of montelukast sodium by capillary electrophoresis”. *Journal of separation science*, vol. 31, pp. 1137-1143, 2008.

[5] I. Alsarra, M. Al-Omar, E. A. Gadkariem and F. Belal. “Voltammetric determination of montelukast sodium in dosage forms and human plasma”. *Il Farmaco*, vol. 60, pp. 563-567, 2005.

[6] H. Ochiai, N. Uchiyama, T. Takano, K. I. Hara and T. Kamei. “Determination of montelukast sodium in human plasma by column-switching high-performance liquid chromatography with fluorescence detection”. *Journal of Chromatography B: Biomedical Sciences and Applications*, vol. 713, pp. 409-414, 1998.

[7] A. K. Shakya, T. A. Arafat, N. M. Hakooz, A. N. Abuawwad, H. Al-Hroub and M. Melhim. “High-performance liquid chromatographic determination of montelukast sodium in human plasma: Application to bioequivalence study”. *Acta Chromatographica*, vol. 26, pp. 457-472, 2014.

[8] S. S. Patil, S. Atul, S. Bavaskar, S. N. Mandrupkar, P. N. Dhabale and B. S. Kuchekar. “Development and statistical validation of spectrophatometry method for estimation of Montelukast in bulk and tablet dosage form”. *Journal of Pharmacy Research*, vol. 2, pp. 714-716, 2009.

[9] M. Arayne, N. Sultana and F. Hussain. “Spectrophotometric method for quantitative determination of montelukast in bulk, pharmaceutical formulations and human serum”. *Journal of analytical Chemistry*, vol. 64, pp. 690-695, 2009.

[10] P. V. Adsule, K. Sisodiya, A. G. Swami, V. P. Choudhari and B. S. Kuchekar. “Development and validation of UV spectrophotometric methods for estimation of montelukast sodium in bulk and pharmaceutical formulation”. *Int J Pharm Sci Rev Res*, vol. 12, pp. 106-8, 2012.

[11] W. Badulla and G. Arli. “Comparative study for direct evaluation of montelukast sodium in tablet dosage form by multiple analytical methodologies”. *Rev Roum Chim*, vol. 62, pp. 173-179, 2017.

[12] S. Muralidharan, L. J. Qi, L. T. Yi, N. Kaur, S. Parasuraman, J. Kumar and P. V. Raj. “Newly developed and validated method of montelukast sodium estimation in tablet dosage form by ultraviolet spectroscopy and reverse phase-high performance liquid chromatography”. *PTB Reports*, vol 2, pp. 27-30, 2016.

[13] K. Singh, P. Bagga, P. Shakya, A. Kumar, M. Khalid, J. Akhtar and M. Arif. “Validated UV spectroscopic method for estimation of montelukast sodium”. *IJPSR*, vol. 6, pp. 4728-4732, 2015.

[14] S. R. Bhagade. “Spectrophotometric estimation of montelukast from bulk drug and tablet dosage form”. *International Journal of Pharmaceutical Sciences and Research*, Vol. 4, pp. 4432, 2013.

[15] K. Pallavi and S. Babu. “Validated UV spectroscopic method for estimation of montelukast sodium from bulk and tablet formulations”. *International Journal of Advances in Pharmacy, Biology and Chemistry*, vol. 1, pp. 450-453, 2012.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Offline Handwritten English Alphabet Recognition (OHEAR)

Hamsa D. Majeed, Goran Saman Nariman

Department of Information Technology, College of Science and Technology, University of Human Development, Kurdistan Region, Iraq

ABSTRACT

In most pattern recognition models, the accuracy of the recognition plays a major role in the efficiency of those models. The feature extraction phase aims to sum up most of the details and findings contained in those patterns to be informational and non-redundant in a way that is sufficient to fen to the used classifier of that model and facilitate the subsequent learning process. This work proposes a highly accurate offline handwritten English alphabet (OHEAR) model for recognizing through efficiently extracting the most informative features from constructed self-collected dataset through three main phases: Pre-processing, features extraction, and classification. The features extraction is the core phase of OHEAR based on combining both statistical and structural features of the certain alphabet sample image. In fact, four feature extraction portions, this work has utilized, are tracking adjoin pixels, chain of redundancy, scaled-occupancy-rate chain, and density feature. The feature set of 27 elements is constructed to be provided to the multi-class support vector machine (MSVM) for the process of classification. The OHEAR resultant revealed an accuracy recognition of 98.4%.

Index Terms: Alphabet Recognition, Handwriting Recognition, Multi-Class Support Vector Machine, Feature Extraction, Optical Character Recognition

1. INTRODUCTION

In the digital world, handwriting is one of the most appeared challenges faced in daily life. When handwriting is detected and transformed into a digital device, several pattern analysis problems will appear that need to be solved. The problems include handwriting recognition, script identification and recognition, signature verification, and writer identification. One of the most challenging and researchable fields among mentioned problems is handwriting recognition. The well-known system in this field is Optical Character Recognition (OCR) which transforms the uneditable text-image format of script into a machine-editable and manageable format of the script. In other words, OCR is a converter software of scanned scripts to a format that could be processed as a character by a computer. For the 1st time, OCR was invented by Carley in 1870 for processing scanned retina [1].

It is worth mentioning that nearly all of the OCR systems are script specific in the sight that they are restricted to recognizing a particular language or a writing system excluding several works that focused on multilinguistic handwriting recognition. However, most works focus on a specific script or language, but still, it has been broken down for more specificity which only covers special symbols, numerals, or characters within the same language or script.

After performed an in-depth review of several research articles including survey articles [2]–[4], we conclude that the entire process of alphabetic handwriting recognition could be classified under some separated classification types based on several factors as below.

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp29-39 **E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Hamsa D. Majeed, Department of Information Technology, College of Science and Technology, University of Human Development, Kurdistan Region, Iraq. E-mail: hamsa.al-rubaie@uhd.edu.iq

Received: 24-05-2022 Accepted: 15-08-2022 Published: 20-08-2022

1. Script writing system
2. Data acquisition (input modes) (online and offline)
3. Granularity level of documents
4. Source of the collected dataset
5. Script recognition process

The scriptwriting system type defines the selected language to be recognized in the proposed system. The languages which are in use today throughout the world have been defined under several different systems, more details can be found in Sinwar *et al.* [2], Ghosh and Shivaprasad [3], Pal [4], Ubul *et al.* [5].

The mechanism of data acquisition could be separated into two categories [2], [6], [7]: Offline and online handwriting recognition. In online handwriting recognition, a digital device with a touch screen without a keyboard must be involved like a personal digital assistant (PDA) or mobile. Where screen sensors receive the switching of pushing and releasing the pen on the screen together with the pen tip movements over the screen. While in offline mode, image processing is involved by converting an input image (from a scanner or a camera) of text to character code which is aimed to be utilized by a text processing application.

Granularity level of documents describes the stage of detailed information taken as initial input to the defined and proposed framework, as example, a full page or a single letter of text image uses as initial input.

There are two types of sources of collected dataset; public dataset (real-world dataset) and self-constructed dataset. The term "public dataset" refers to a dataset that has been saved in the cloud and made open to the public. MNIST, Keras, Kaggle, and others are examples. While the self-constructed dataset is the dataset that the researchers create and prepare on their own by scanning handwritten documents from different people.

The script recognition process is the primary section which is the practical part of the work. In general, it is formed from four main phases, namely, preprocessing (P), segmentation (S), feature extraction (F), and classification (C). The last two phases, F and C, are the common phases in the study, there is not any work without any of these two phases. However, there are many researches in literatures without P and/or S.

2. RELATED WORK

In this section, several works will be illustrated in the field of English alphabet handwritten recognition for bringing to light varied methodologies employed in each step to accomplish the recognition.

Starting with a review study [8] which summarizes eight research papers with their contributions, limitations coupled with strategies employed to enhance OCR systems. Here, we mention two of them and demonstrate their conclusion; Patel *et al.* [9] was working on the ANN (Artificial Neural Network). Characters were extracted using MATLAB. The module was analyzed pixel by pixel and transformed into a list of characters. To find edges, they used an edge and skew detection algorithm. Moreover, it became normalized thereafter. The authors claim that the accuracy is improved by increasing the hidden layers and neurons. Only 100 input neurons were used for testing which accounts for the work's limitation. The litterateurs of Gupta *et al.* [10] segment the input data at the word level into separated characters using AI and heuristic functions. Then, the feature vector is generated by extracting features from the segmented characters. As a property of vectors, blending three types of Fourier descriptors are utilized in parallel. Finally, SVM has been employed as a classifier. The authors claim that a piece of recognition error rates may arise from the usage of low-quality material and ink density diversity, as well, is another point that degrades document quality.

The authors of Karthi *et al.* [11] propose a system to recognize cursive handwriting English letters. The initial system input is in pdf format of both alphabet and cursive English letters which have been gathered from 100 different people and the total samples are 2K. This module is accomplished through four processes, namely, image preprocessing, skeletonization, segmentation points identification, and contour separation. The final module utilizes a convolutional neural network (CNN) for training the dataset to predict recognition. Support vector machine (SVM) is the system classifier. The accuracy rate of this work achieved 95.6%.

The investigation of pre-processing, feature extraction, and classifier techniques is emphasized in Ibrahim *et al.* [12]. The pre-processing initiates with normalizing image letters to 70X50 pixel dimensions by utilizing the nearest neighbor technique. Then, the binarization process is executed using Otsu's threshold sampling procedure. Character skeleton and contour algorithms have been employed to accomplish the feature extraction step. Further, both isolated and combined feature extraction procedures are involved in the experiments. The study employed two different classifications (Hibbert

Classifier) techniques which are support vector machine (SVM) and multilayer perceptron (MLP) classifiers. The recognition experiment outcomes obtained an accuracy of 97%.

In Parkhedkar *et al.* [13], a system has been produced that implements all four available steps of the handwritten recognition process. It takes a scanned document as initial input and proceeds through preprocessing for the oncoming step in which each letter of the word will be separated from the other (segmentation). Then, the Gabor feature is served for extraction of the features that will be passed through the KNN classifier on the final step. The accuracy rate of the developed project has not been given. Rather, the authors claim that multiple experiments have been established using publicly available data and the achieved accuracy is the highest in the experimentation studies when using constant data.

In Gautam and Chai [14], the proposed work uses the publicly available dataset EMINST and MINST. This means that no pre-processing and segmentation have been applied. The work only focuses on the last two steps, namely, feature extraction besides classification. The features of English letters and digits have been extracted by employing a hybrid proposal, which combines the zoning method and zig-zag diagonal scan. The feedforward NN (FFNN) is utilized as a classifier. Then, the back-propagation learning algorithm is used for the network training. The accuracy rate of English characters (e EMNIST) and English numbers (MNIST) recognition stands for 99.8% and 94%, respectively.

The litterateurs of Zanwar *et al.* [15] select 3410 samples of Chars74K which is another publicly available dataset. Independent component analysis (ICA) technique is used in feature extraction phase. Backpropagation neural networks have been employed in the final phase (classification). The recognition accuracy shows 98.21% of matching characters.

The authors of the previous study have improved their work [16] by hibernating two techniques at the feature extraction phase while the rest remained the same apart from the dataset that MNIST employed in this work. The new technique integrates detached component analysis and hybrid PSO and firefly optimization for effective selection of features and then applies a supervised learning technique called backpropagation neural network to perform classification. Recognition accuracy scores of 98.25% were recorded using the models.

3. PROPOSED METHOD

The proposed technique for offline handwritten English alphabet recognition (OHEAR) is revealed in this section. According to the aforementioned classification of handwritten recognition, Table 1 shows the used category of the classes for the presented method.

The selected input script to the model is the English alphabet (capital and small). The presented approach acquires data offline, which implies that scanned documents (images) are served as an entry to the model. Because the model operates at the character level, it takes character images as input. The used dataset nature is self-constructed, stating that it was manually gathered from 120 individuals, each of whom typed 52 characters from A to Z and a-z.

The contribution takes place in the general script recognition process phases which are the primary and the heart of such works. Apart from data acquisition which was mentioned before (commonly referred to as the first phase), it is divided into three major phases (PFC), which are pre-processing, feature extraction, along with classification. Each phase's output will be provided into the next. The phases are illustrated in Fig. 1 and described in the subsections that follow.

TABLE 1: Classification of the proposed technique

Classes	Nominated category
Script writing system	English alphabet
Data acquisition	Offline
Granularity level of documents	Character level
Source of the collected dataset	Self-constructed dataset
Script Recognition Process	PFC

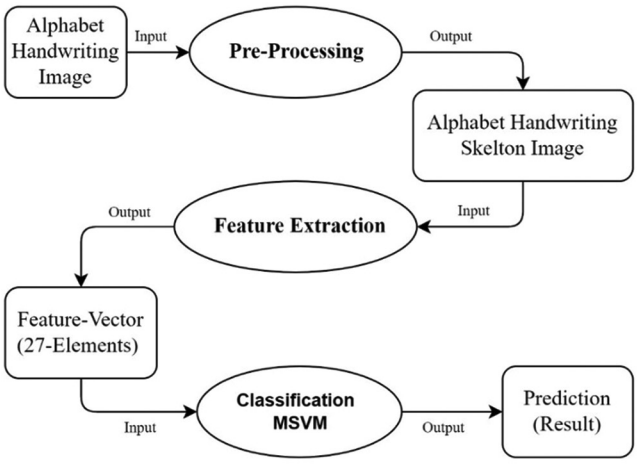


Fig. 1. Script recognition process of the proposed model.

3.1. Pre-Processing

This step is required and is a critical procedure because we are using a self-constructed dataset rather than public datasets. It should be carefully studied because the model's accuracy rate directly leans on the output quality of this phase. The reason being such a dataset used instead of using the small image size, cleaned, and noise-free public dataset is that it's truly close to data actuality in terms of real-world application.

The pre-processing procedure is broken down into six isolated processes, as shown in Fig. 2. The initial process is converting the inputs to grayscale for the purpose of size reduction which implies higher performance for the following processes without affecting accuracy.

The contrast enhancement manipulates and redistributes image pixels to improve the partitioning of hidden structural variations in pixel intensity to assemble a more distinct structural distribution.

The distribution of the pixels is calculated utilizing the histogram equalization (HE) approach, which represents the probability allocation of the image's gray levels (pixels).

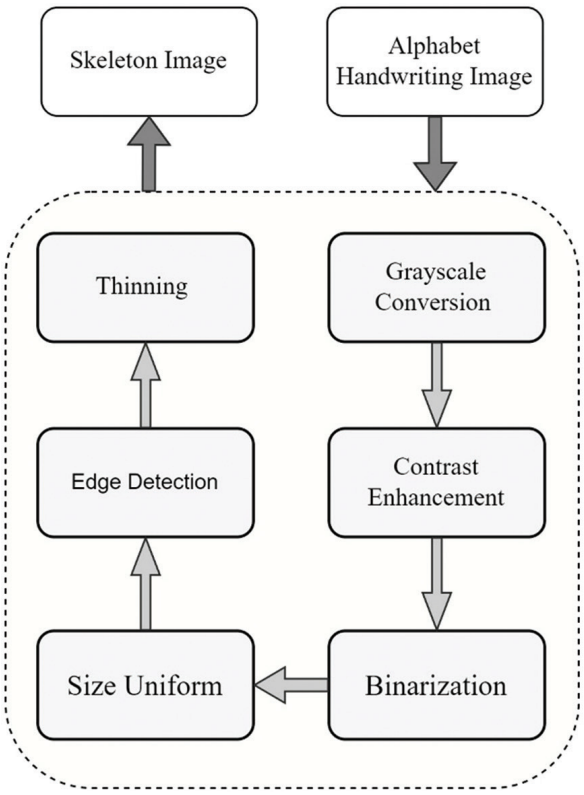


Fig. 2. The pre-processing phase of the proposed approach.

Adaptive thresholding, based on Otsu's approach, was used to convert the grayscale picture to a binary image (Binarization). This technique is used to divide the pixels into two classes: Foreground and background. Following the creation of the binary image, the sizes of all input images are uniform such that the output image only comprises the English letter. The compromised area refers to the region of interest (ROI).

After size uninformed, edge detection is the next step. It was done using the Canny approach, which locates all edges with the shortest distance between the detected edge and the processed letter's true edge. The final step of the pre-processing is for usage of skeletonization and thinning to produce the skeleton of the letter image. The thinning technique removes black foreground pixels, one at a time, until a skeleton of one-pixel width is obtained.

3.2. Feature Extraction

This phase is the uppermost critical and crucial because a proper feature extraction mechanism should be selected for a specified script. It is obvious that various scripts have distinct properties, therefore, factors that are effective in recognizing one script may not be effective in identifying another. The primary contribution of this study is the identification of features of English letter patterns that will be extracted and prepared for the oncoming and final phase of the recognition process. The feature vector is the output that consists of four segments as illustrated in Fig. 3. Each segment of the extracted feature vector is described below:

3.2.1. Tracking Adjoins Pixels

The first step in feature vector creation starts with studying the image details at the pixel level, discovering the starting point then tracking the flowing of each letter through the pixels owned by concerned image. Any pixel with more than 2 adjoins is represented as an intersection point, while the open-end point has precisely one adjoin as illustrated in Fig. 4 which is the English Letter H with two intersection points and four open-ended points.

3.2.2. Chain of Redundancy (CR)

The next feature is retrieved using Freeman Chain Code [17]. In the proposed OHEAR, the Chain code is employed to describe the form of English alphabets as a linked sequence of pixels in a restricted length and direction. This expression is based on clockwise 8-connectivity, as shown in Fig. 5a.

The skeleton image is tracked starting from the open-ended pixels and stopped at the last open-ended pixel. As for the intersection pixel, the tracking operation will be done

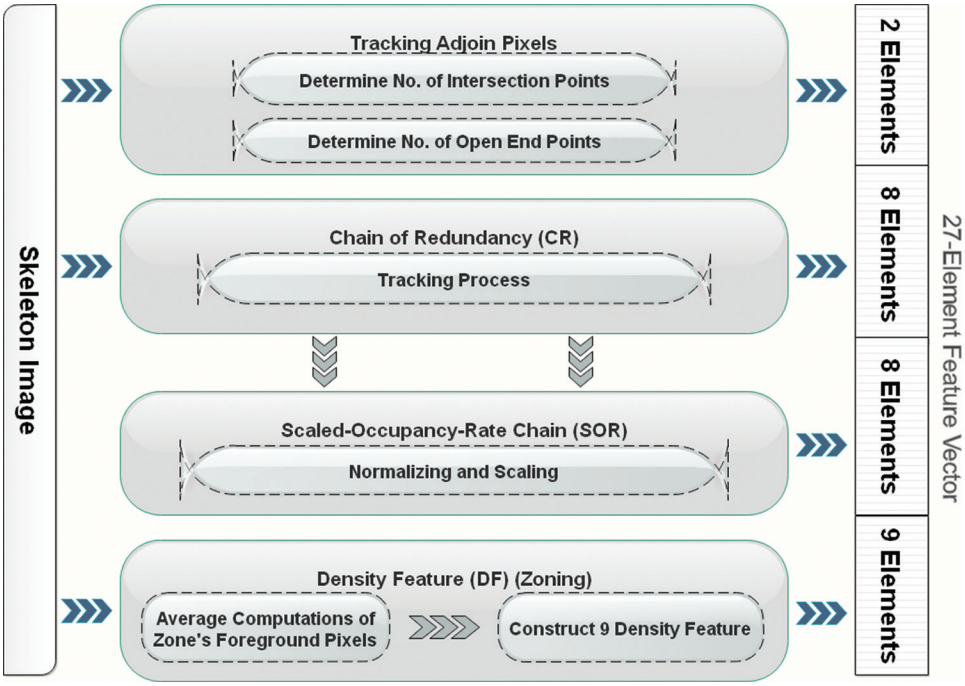


Fig. 3. Feature extracted process of OHEAR.

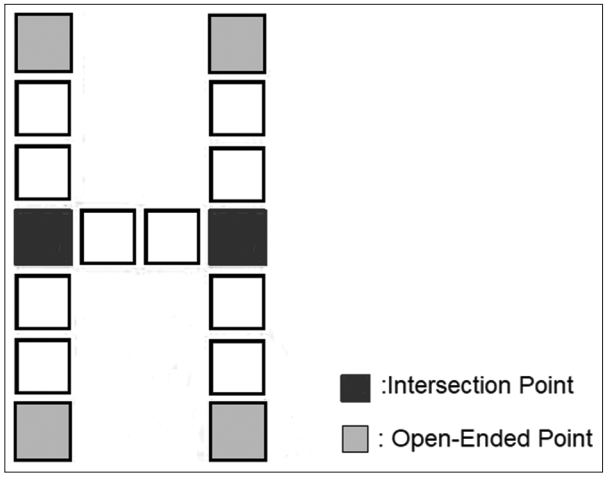


Fig. 4. Intersection points and open ended of letter H.

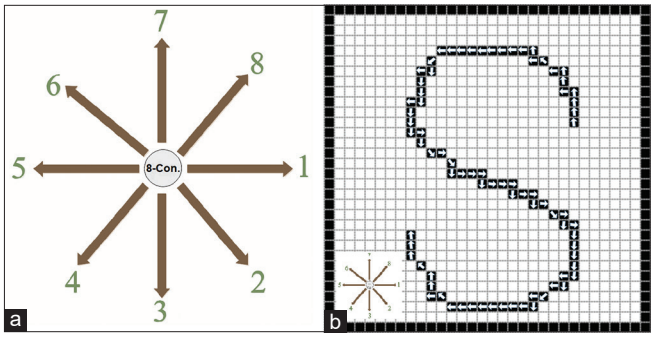


Fig. 5. (a) Eight directions of freeman chain code. (b) S letter with chain code directions.

by proceeding in the alternative direction defined by that intersection point until it gets to the terminated open-end pixel. This process will continue until the entire pixels of the entered skeleton image of the English alphabet are tracked. A numbering method is employed to code the direction and length belonging to the pixels.

For instance, the generated chain code for the letter (S) is illustrated in Fig. 5b which shows that the starting pixel is the top-right open-ended one which indicates chain code 7 followed by three more 7s. Then, it turns to the left as 5 and so on. These chain code numbers will be adjusted for creating the Change of Redundancy (CR). CR consists of eight elements starting from index 1 to 8 which index numbers represent the directional numbers from the freeman chain code. For instance, in the full tracking process, 11, 3, and 19 times the chain code directions of 1, 2, and 3 have been repeated, respectively. In the result, the indexes 1, 2, and 3 of CR contain 11, 3, and 19. Finally, the CR with eight elements will be added to the feature vector as the second segment.

3.2.3. Scaled-Occupancy-Rate chain (SOR)

More valuable information can be retrieved from the above-generated data (CR) which involves the total pixels' number occupied by the English letter and considering the repetition

of the individual number chain code directions. The Scaled-Occupancy-Rate chain (SOR) can generate a reasonable value to be added to the feature vector that could be generated, the Scaled-Occupancy-Rate chain (SOR).

SOR is a significant segment of the feature vector that gives weight to each chain code direction. For instance, the ideal CR of direction 3 (from Fig. 5a) for letters I and E is similar but the SOR of them is totally different, it gives 100% weight to the direction of 3 for I but much less for E.

SOR will be generated as follows, the division process applied to each index of CR on the total pixels number occupied by the skeleton image of the English letter, in other words, each index of CR is divided by the summation of CR's indexes values. For instance, from mentioned CR of S, the total pixels number of S's foreground is 76, so, the computation of the first and third indexes will be 11/76=0.144 and 19/76=0.25, respectively.

Finally, a scale factor of 10 will be hands-on to get a more practical value for classification objectives. For example, 0.144 and 0.25 will be 1.4 and 2.5, respectively. The final result with eight elements will be added to the feature vector as the third segment.

3.2.4. Density Feature (DF)

The final insertion to the feature set is the information extracted from the demanded character under the employment of the density feature. This segment of feature is achieved using the zoning technique which has been applied to the skeleton image of letters.

Zoning is a statistical feature extraction that calculates the density of foreground pixels by the zone's pixel numbers, each letter's image divided into 9 (3 × 3) zones. The zone's size of each is 10 × 10 denoting that the entered image will be resized to 90 × 90 before these divisions are applied as illustrated in Fig. 6.

This density feature (DF) will be calculated for all nine zones. Consequently, nine values will be generated and will be added to the feature vector as the last segment.

The ideal (S) illustrated in Fig. 6 takes all the nine zones, but in reality, the handwritten is dissimilar from the ideal state, the results from our dataset plotted in Fig. 7 demonstrate that with different handwritten styles, the zones' occupation will be changed accordingly. Fig. 7A3 shows that zone-3 and zone-7 will be discarded in the calculation of DF because they have zero density. It is alike the situation for Fig. 7B3 zone-3.

3.3. Classification

The latest phase of the proposed approach is the classification process which determines the recognition output of the given English letter's image. The multi-class support vector machine (MSVM) has been implied which is based on the support vector machine (SVM) technique.

The SVM is a well-known classifier, and it has obtained much traction in machine learning and statistics since it was first introduced. Vapnik's foundational work (1998) [18] set the groundwork for the theory of SVM generic statistical

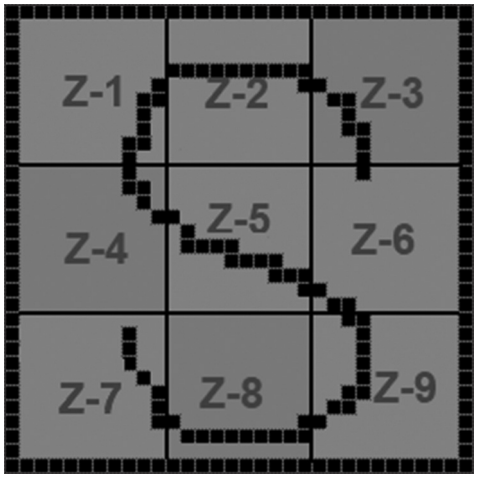


Fig. 6. Ideal resized and zoned S letter.

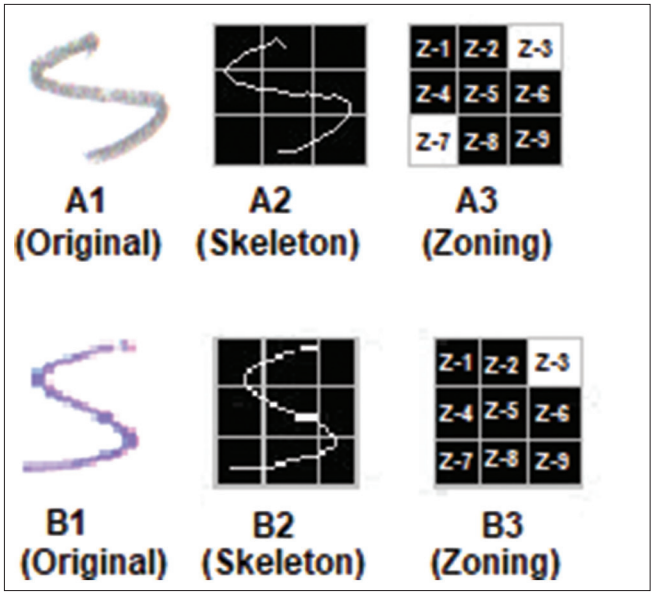


Fig. 7. Zone density and occupation of different handwritten styles.

learning, which, in turn, inspired several expansions. SVM is a binary classifier which means it only handles two-class classification issues. Therefore, it does not suit our work while having 52 English alphabet classes. More details about binary SVM can be found in Cristianini and Shawe-Taylor [19], Schoelkopf and Smola [20]. As a result of its limitation, the MSVM model has been developed to determine the dynamic process instability using multi-class classification. It has also found use in a variety of fields, including control chart pattern recognition besides industrial problem diagnosis [21], and is employed for many different language characters and numerals recognition such as Romaine, Thai, French, and Arabic Persian [1]. Furthermore, it is worth noting that, according to Ubul *et al.* [5], MSVM classifiers using various extracted features outperformed K-NN and NN classifiers in handwritten recognition field.

The feature vectors from the previous phase which were generated from 80% of the self-constructed dataset will be employed to train MSVM to create the classification model. This model creates 52 classes of small and capital English letters. The remaining 20% dataset are for testing operation.

4. RESULTS

The experimental outcomes have been established to assess the proposed model OHEAR performance. The model is implemented using MATLAB 2020a and the evaluation process had been performed through a constructed dataset consisting of 52 offline handwritten English alphabet from A (a) -to -Z (z) self-collected from 120 individuals, in a total of 6240 samples collected for capital and small letters together. With the aim of covering most of the various possibilities of the handwritten patterns, various types of writing objects (pen, pencil, and magic marker) with different colors and font sizes were applied to prove the effectiveness of the presented model regarding the recognition process.

The first set of results was in image form and from the share of preprocessing phase, as Fig. 2 presents, this phase goes through six stages starting from greyscale conversion to thinning, the outcome of this phase is illustrated in Fig. 8 for letter G.

Regardless of the entered image's color, it will be converted into the grayscale in the early steps of preprocessing phase, in the second stage, the brightness level is equalized yielding the contrast enhancement of that image. The oncoming stage shows the outcome of binary conversion through the

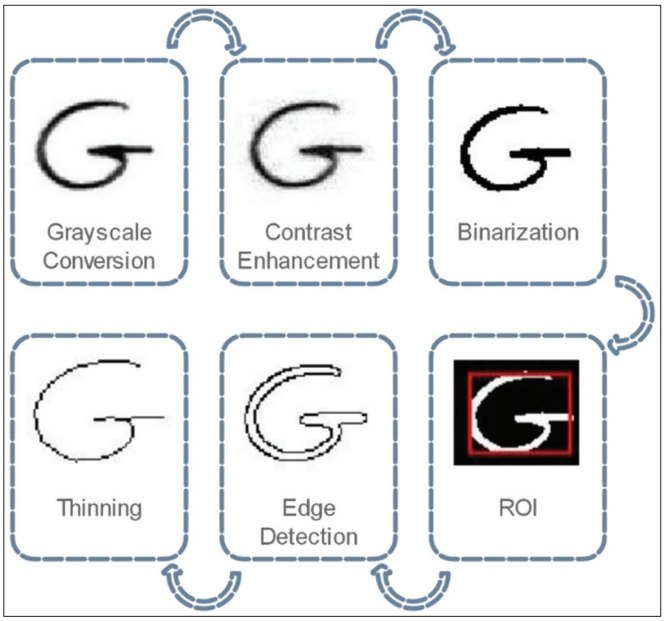


Fig. 8. Pre-processing phase results.

adaptive thresholding of the input. Size refinement is applied after binarization to determine ROI in a preparation step to the following stage where the edge of the interesting region is detected, the final stage represents the resultant thinning output to be ready for the oncoming phase of OHEAR which is the feature extraction. In this phase, the same stages are applied for all the 52 letters for each individual. In fact, it applied to all the collected dataset to the feature extraction state.

The second set of results was in the numbers form where the feature set has been extracted for each letter of the English alphabet through the OHEAR model where the statistical and structural features have been extracted and combined into one feature set with 27 elements.

As Fig. 3 revealed, the 27 elements of the extracted features are combined in four portions. In this section, those elements are translated into numbers in four tables, each table describes one of those portions in different capital with small letters. In fact, each table describes the outcomes of that portion of the feature set for all of the 52 letters, but for the publication requirements, the letters are distributed among the tables to show most of the outcomes of the letters. Consequently, all portions were gathered from the tables to define each letter in the dataset so as to create the feature set of that letter to be distinguished by the used classifier.

Table 2 illustrates the first portion of the feature set which is outlined by the pair (intersection and endpoints), the

TABLE 2: Intersections and endpoints

Character	A	a	B	b	C	c	D	d	E	e
No. of intersection points	2	2	1	1	2	2	0	2	1	0
No. of endpoints	2	1	0	1	0	0	0	2	3	1

TABLE 3: Chain of redundancy (CR)

Character	Chain of redundancy (CR)							
F	4	4	6	3	13	0	0	1
f	5	3	16	2	1	0	0	2
G	18	5	9	20	12	5	2	6
g	4	7	19	5	8	2	3	4
H	12	3	20	1	0	0	0	1
h	0	5	25	1	0	0	0	0
I	10	7	23	4	4	1	0	0
i	0	3	19	1	0	0	0	0
J	0	1	20	10	6	2	5	5
j	0	2	14	4	4	2	0	0

outcomes of A-to-E small and capital letters were illustrated, some challenges appear in this section of feature collection one of belonged long to the handwriting style in which the lines were not connected properly or more intersection points than normal created. Hence, this portion alone was not reliable enough and needs to have more features to be extracted, which lead to the second and third portions where their results are illustrated in Tables 3 and 4.

Tables 3 and 4 contain the portions: Chain of redundancy (CR) and Scaled-Occupancy-Rate chain, respectively, each portion has eight elements. The outcomes of F-to-J small and capital letters were illustrated in Table 2, while Table 3 shows the outcomes of K, L, M, Y, and P small and capital letters, the letters in Table 3 are not consecutive as trying to decrease the letters with a looks like letters as capital and small or looks as other letters in the same table. Those two portions increased the richness of the extracted characteristics from the letters with a minimum number of feature elements. Moreover, the combination of features' outcomes of the three tables so far improved the classification accuracy. Yet, some limitations floating to the surface of the process, because of existing different techniques in handwriting tracking the chain through the directions may differ for the same letter, for example, the straight line in a letter been written in bent way, or circles in some letter were not written completed, otherwise, some handwritten styles write circles where it should be a normal line, all these issues affect the chain creating process in those portions because it leans on the directions. These limitations have been solved by using another portion of combination which is the Density Feature.

Reaching Table 5 which reports the last portion of the feature set, the density chain provides the feature vector with the last nine elements. Those elements describe the density of nine zones for each letter, the results of Q, R, N, T, and U letters capital and small. Combing the outcomes of this portion with the previous chains boost the recognition accuracy, it gives occupied zones for each letter with the exact rate of that occupation in each zone, which advances the amount of information that extracted about each letter although there are some issues appear in some letters causing due to the writing direction sometimes it's in slant or diagonal way but when it's combined with the other features from the other portions it gives a cleared version of description to the classifier for recognition operation of that letter.

The next and final phase in the OHEAR model is classification, multi-class SVM is employed for this purpose in the proffered model, as a preparation step for this phase, all the features are gathered from the collected samples and then grouped into two packs of data, training data which contain 80% of the constructed dataset (96 samples out of 120 for each letter) fed to classifier for the purpose of training, while the remaining 20% labeled as test data (24 samples out of 120 for each letter) supplied to classifier for performance testing of the presented recognition model.

The recognition accuracy out of 100% has been measured for all the gathered samples. According to the outcomes from the self-constructed dataset used in this study, the handwritten English alphabet recognition accuracy in the proposed model can be classified into three groups:

First Group: The letters which achieved 100% accuracy throughout all the testes samples regardless of the font size, type of used pen, or its color, accompanied by the variety in how it's written or how straight it is (mostly slanted). The proposed combination of feature extraction mechanisms powered up the recognition ability of the classifier. Most of the letters (capital and small) belong to this group and this matter caused the raise of the total recognition accuracy of the proposed model.

Second Group: Portion of the letters which belong to this group, precisely (small letter of L, Capital letter of I, and z) are not fully recognized successfully, the classifier misclassifies one sample from the testing set of samples (i.e., 23 from 24 testing sample scored). This is due to the common way of handwriting those letters, commonly capital letter of I is similarly written as a small letter of L, beside the used way of writing the capital letter of Z with an extra line in the middle which confused the first portion of the feature vector.

Third Group: The letters (i and j) are the reason for this group creation, the classifier misclassifies two of the testing samples (i.e., scored 22 out of 24) for two major reasons, first, the dot (.) above the letters sometimes writing close to the letter, far, or lightly written in a way that excluded in the preprocessing phase. The second reason is produced by ROI determination, when the dot is written far from the letter, then it is considered out of the region of interest and excluded from the process.

Despite the fact that the model achieved an excellent recognition rate of (98.4%), there are still areas for improvement, such as reconsidering the mentioned issues in classification groups, which will be discussed in the following section.

The proposed combination of extracted features in this work is unique, for that matter, a comparison study has been made for the percentage of recognition rate achieved by other researchers that used different approaches for feature extraction as Table 6 illustrates. It is noticeable that the proposed model contributes remarkable efficient

TABLE 4: Scaled-occupancy-rate chain (SOR)

Character	Scaled-Occupancy-Rate chain (SOR)							
K	0.1875	0.3437	0.0312	0.0343	0.0937	0	0	0
k	0.2285	0.4571	0.3142	0	0	0	0	0
L	0.2142	0.0714	0.4285	0.2142	0	0	0	0.0714
l	0	0.1666	0.8333	0	0	0	0	0
M	0.0476	0.1309	0.4047	0	0	0	0.2023	0.2142
m	0.1904	0.0714	0.2857	0.0714	0	0	0.0952	0.2857
Y	0	0	0.2500	0.7250	0.0250	0	0	0
y	0.0212	0.1276	0.4042	0.2127	0	0	0	0.1702
P	0	0.0555	0.6944	0.2222	0	0	0	0
p	0.0681	0.0681	0.2045	0.1136	0.0101	0	0	0

TABLE 5: Density features

Character	Zones density values								
	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9
Q	14.166	15.111	12.277	14.166	15.111	16.055	10.622	22.133	0
q	10.818	17	11.333	17	27.818	5.6666	0	0	12.750
R	34.151	36.428	0	31.875	30.222	11.333	19.125	0	20.777
r	21.250	7.0833	18.888	21.250	28.333	0	17.163	4.3589	0
N	3.2692	8.1730	16.346	19.615	19.615	19.615	15.088	15.088	9.0532
n	13.909	23.181	4.2148	6.9545	23.181	23.181	0	18.545	6.3223
T	16.071	28.928	13.928	0	15	0	0	12	0
t	0	10.699	5.3496	13.730	35.664	10.699	0	23.181	12.482
U	20.863	0	11.590	25.500	0	23.181	11.590	23.181	9.2727
u	16.227	0	8.4297	25.500	9.2727	23.181	2.3181	13.909	14.752

TABLE 6: Illustrations of accuracy rates for various feature extraction techniques

Previous work	Feature extraction approach	Accuracy rate
Gautam and Chai [14]	Combination: Zoning method+zig-zag diagonal scan	94%
Zanwar *et al.* [16]	Combination: Detached component analysis+hybrid PSO	98.25%
Ibrahim *et al.* [12]	Combination: Features that are based on viewing capabilities+bit map feature.	97%
Zanwar *et al.* [15]	Independent component analysis (ICA) technique	98.21%
The proposed model (OHEAR)	Combination: Tracking adjoin pixels+chain of Redundancy+Scaled-Occupancy-Rate chain+and density feature	98.4%

recognition performance with a non-previously processed self-constructed dataset with different types of writing objects along with avoiding redundancy in the generated data for classification purposes.

5. CONCLUSION AND FUTURE CONSIDERATION

The most compacted and informative set of features has remarkable effectiveness to enhance the classifier – efficiency, recognition accuracy, and reliable classification accomplishment. This work presents an optimized feature extraction phase by employing both statistical and structural techniques to retrieve the features from constructed dataset self-collected for offline handwritten English alphabets through recognition (OHEAR) model. The extraction process goes through four stages: Tracking adjoins pixels, redundancy chain, adjusted scaled redundancy chain, and density feature.

The extracted feature set is provided to the multi-class SVM classifier which has been trained and tested using 120 sets of each capital and small letters of handwritten English alphabets. The proffered model achieved a recognition accuracy of 98.4%. Despite the good recognition rate, the experimental outcomes reveal some misclassification of some letters, those issues could be enhanced by making slight changing in the used features extraction techniques to raise the classification accuracy. Replacing the tracking adjoin pixels with another technique is a suggestion to overcome those misclassification issues, adopting the actual length of chain before redundancy calculation as a number in the features set are another possible suggestion besides expanding the threshold of ROI to include all the detailed characteristics of the letters while still, the increasing of the training set is always a valid option to improve the classification accuracy process. All over, reducing the total length of the feature vector with preserving the quality of the system and the level of validation rate is the goal looking forward to, on the other hand, employing another classifier is an important factor to achieve an optimum outcome from the proposed system. Moreover, the presented recognition model (OHEAR) can be extended for symbols, special characters, or other language recognition.

REFERENCES

[1] J. Mantas. "An overview of character recognition methodologies". *Pattern Recognition*, vol. 19, no. 6, pp. 425-430, 1986.

[2] D. Sinwar, V. S. Dhaka, N. Pradhan and S. Pandey. "Offline script recognition from handwritten and printed multilingual documents: A survey". *International Journal on Document Analysis and Recognition*, vol. 24, no. 1-2, pp. 97-121, 2021.

[3] D. Ghosh and A. P. Shivaprasad. "Handwritten script identification using the possibilistic approach for cluster analysis". *Journal of the Indian Institute of Science*, vol. 80,no. 3, pp. 215, 2000.

[4] U. Pal. "Automatic script identification: A survey". *J. VIVEK, Bombay*, vol. 16, no. 3, pp. 2635, 2006.

[5] K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo and I. Yibulayin. "Script Identification of Multi-Script Documents: A Survey". *IEEE Access*, vol. 5, pp. 6546–6559, 2017.

[6] A. Priya, S. Mishra, S. Raj, S. Mandal and S. Datta. "Online and offline character recognition: A survey". *2016 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0967-0970, 2016.

[7] X. Y. Zhang, Y. Bengio and C. L. Liu. "Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark". *Pattern Recognition*, vol. 61, pp. 348-360, 2017.

[8] B. M. Vinjit, M. K. Bhojak, S. Kumar and G. Chalak. "A Review on Handwritten Character Recognition Methods and Techniques". In: *2020 International Conference on Communication and Signal Processing* (*ICCSP*), 2020

[9] C. I. Patel, R. Patel and P. Patel. "Handwritten character recognition using neural network," *International Journal of Scientific and Engineering Research*. vol. 2, no. 5, pp. 1-6, 2011.

[10] A. Gupta, M. Srivastava and C. Mahanta. "Offline handwritten character recognition using neural network". In: *2011 IEEE International Conference on Computer Applications and Industrial Electronics* (*ICCAIE*), 2011.

[11] M. Karthi, R. Priscilla and K. S. Jafer. "A novel content detection approach for handwritten English letters". *Procedia Computer Science*. vol. 172, pp. 1016-1025, 2020.

[12] B. Ibrahim, H. Yaseen and R. Sarhan. "English character recognition system using hybrid classifier based on MLP AND *SVM". International Journal of Inventions in Engineering and Science Technology*, vol. 5, pp. 1-15, 2019.

[13] S. Parkhedkar, S. Vairagade, V. Sakharkar, B. Khurpe, A. Pikalmunde, A. Meshram and R. Jambhulkar. "Handwritten English character recognition and translate English to Devnagari words". *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, pp. 142-151, 2019.

[14] N. Gautam and S. S. Chai. "Zig-zag diagonal and ANN for English character recognition". *International Journal of Advanced Research in Computer Science*. vol. 8, no. 1-4, pp. 57-62, 2019.

[15] S. R. Zanwar, A. S. Narote and S. P. Narote. "English character recognition using robust back propagation neural network". In: *Communications in Computer and Information Science*. Springer, Singapore, pp. 216-227, 2019.

[16] S. R. Zanwar, U. B. Shinde, A. S. Narote and S. P. Narote. "Handwritten English character recognition using swarm intelligence and neural network". In: *Intelligent Systems, Technologies and Applications*. Springer, Singapore, pp. 93-102, 2020.

[17] H. Freeman. "Computer processing of line-drawing images". *ACM Computing Surveys*, vol. 6, no. 1, pp. 57-97, 1974.

[18] V. N. Vapnik and V. Vapnik. "*Statistical Learning Theory*". Vol. 1. Wiley, New York, 1998.

[19] N. Cristianini and J. Shawe-Taylor. "Background Mathematics".

In: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, pp. 165-172, 2013.

[20] B. Schoelkopf and A. J. Smola. "Learning with kernels: Support vector machines, regularization, optimization, and beyond," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 16, no. 3, pp. 781-781, 2005.

[21] Y. Liu and H. Zhou. "MSVM recognition model for dynamic process abnormal pattern based on multi-kernel functions". *Journal of Systems Science and Information*, vol. 2, no. 5, pp. 473-480, 2014.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Plate Number Recognition based on Hybrid Techniques

Hemin Omer Latif[1], Hawar Hussein Yaba[2]

[1]Department of IT/SE, American University of Iraq, Sulaimani, Sulaymaniyah, Iraq, [2]Computer Institute in Sulaymaniyah, Sulaymaniyah, Iraq

ABSTRACT

Globally and locally, the number of vehicles is on the rise. It is becoming more and more challenging for authorities to track down specific vehicles. Automatic License Plate Recognition becomes an addition to transportation systems automation. Where the extraction of the vehicle license plate is done without human intervention. Identifying the precise place of a vehicle through its license plate number from moving images of the vehicle image is among the crucial activities for vehicle plate discovery systems. Artificial intelligence systems are connecting the gap between the physical world and digital world of automatic license plate detection. The proposed research uses machine learning to recognizing Arabic license plate numbers. An image of the vehicle number plate is captured and the detection is done by image processing, character segmentation which locates Arabic numeric characters on a number plate. The system recognizes the license plate number area and extracts the plate area from the vehicle image. The background color of the number plate identifies the vehicle types: (1) White color for private vehicle; (2) red color for bus and taxi; (3) blue color for governmental vehicle; (4) yellow color for trucks, tractors, and cranes; (5) black color for temporary license; and (6) green color for army. The recognition of Arabic numbers from license plates is achieved by two methods as (1) Google Tesseract OCR based recognition and (2) Machine Learning-based training and testing Arabic number character as K-nearest neighbors (kNN). The system has been tested on 90 images downloaded from the internet and captured from CCTV. Empirical outcomes show that the proposed system finds plate numbers as well as recognizes background color and Arabic number characters successfully. The overall success rates of plate localization and background color detection have been done. The overall success rate of plate localization and background color detection is 97.78%, and Arabic number detection in OCR is 45.56 % as well as in KNN is 92.22%.

Index Terms: Image, Automatic Number Plate Recognition, Background-Color detection, Arabic Number Recognition, K-Nearest Neighbors, Tesseract-OCR

1. INTRODUCTION

1.1. Overview

The large challenge in several nations where road traffic and additionally traffic-related crime is increasing day by day such as swiping hit and also run kidnapping murder despoilment as well as additionally trafficking. With the improvement of people's living standards, every household now has a vehicle [1]. In contemporary life, we have to encounter numerous problems, one of which is traffic congestion ending up being severe extra day after day. It is necessary to adopt the relevant, intelligent traffic management system to strengthen the city security prevention and control and the effective management of motor vehicles in the city management. Nowadays, in the automobile era, the license plate recognition system has become the inevitable product of intelligent human life. Its application scope is vast: The management of vehicle access in residential areas, the detection of speeding vehicles on highways, parking lot

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp39-48 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Hawar Hussein Yaba, Computer Institute in Sulaymaniyah, Sulaymaniyah, Iraq.
E-mail: hawar.yaba@spu.edu.iq

Received: 21-01-2022 Accepted: 22-06-2022 Published: 01-09-2022

management, and so on need the assistance of the license plate recognition system. It is a crucial place of a research study because of its applications like car parking, availability control, roadway tolling, and also authorities. The research and development of the Intelligent Transportation System (ITS) [2] supply information about vehicle identification using plate numbers, which can be utilized in evaluations and control.

Automatic License Plate Recognition [3] is a real-time machine-intelligent as well as the ingrained system which recognizes the vehicles straight from the photo of the number plate. Because of modern technology and additionally the raise in the use of vehicles, the requirement for a machine-oriented acknowledgment along with checking system is of immense significance. Each vehicle on the planet should have its really own number plate that installed on its body (a minimum of on the back). Each vehicle has serial number perception on a plate; therefore, there is no need for external cards, tags, or transmitters. The vehicle identifying permit plate system replaces the manual permit plate number composing process in the computer system. ANPR system has been extensively taken on for number plate discovery for the English Language in industrialized nations. In this paper, a method exists for an automatic license plate recognition system that deals with the difficulties of Arabic permit plate system faced presently. The number plate of vehicle detection is executed for Iraq locations using the Arabic Language. The project fundamentals representation in number 1 revealed over.

The license plate identification technique consists of three main topics. (1) Find number plate area from vehicle images. (2) Identify the background color of the license plate. (3) Arabic numeric character segmentation and (4) Arabic numeric character recognition. The number plate background color [4] identifies the vehicle types: (1) White color for private vehicle; (2) red color for bus and taxi; (3) blue color for governmental; (4) yellow color for trucks, tractors, and cranes; (5) black color for temporary license; and (6) green color for army vehicles. Arabic Number Recognition is carried out using two methods. (1) KNN Machine Learning Algorithm and (2) Tesseract OCR.

1.2. Problem Background and Research Objective and Scope

To identify the automated license plate from the image captured, we need to get the precise location of the plate. Image segmentation plays a fundamental duty, so regarding finding a plate from the image. Different background color plates show different categories of vehicle types. Accurate color detection is a challenging task due to different variable appearance and ranges. There are many methods for character recognition of license plate images [5], such as template matching method, pattern recognition algorithm, artificial neural network recognition algorithm, structural feature recognition algorithm, and statistical feature recognition algorithm. Finding the best practices for plate extraction, background-color detection, and number recognition are a challenging task.

License plate discovery, as well as recognition, is among the significant facets of using the image processing techniques toward smart transportation systems. It is critical to deal with the difficulties of impeding traffic as well as public safety and security renovation in the Kurdistan Regional of Iraq (KRI). In Iraq, the licensed vehicle has six types of plates with different backgrounds and Arabic font color. The background color of the number plate identifies the vehicle types. We first need to set up a number template. The 0–9 in Arabic numbers representing ٠٩٨٧٦٥٤٣٢١license numbers. Arabic numbers get recognized through artificial intelligence through various algorithms of Machine Learning and Optical Character Recognition (OCR). The challenging task is to find the best algorithm with the highest accuracy that recognizes background color and Arabic numbers from the extracted vehicle plate.

License plate detection locates the accurate area in the image as an essential step of the license plate recognition system. First, the research goal aims at plate extraction from a licensed vehicle image correctly. The plate's background colors are different for identifying different vehicles. Second, it aims at detecting background color efficiently for vehicle type recognition. The project implements two main techniques for Arabic Number Recognition as follow: (1) Training Testing based K-mean Nearest Neighborhood Supervised Machine Learning (KNN) and (2) Tesseract Optical Character Recognition (OCR) that aims to find the best technique among them to yield an overall accuracy.

2. LITERATURE REVIEW AND BACKGROUND STUDY

2.1. Vehicle License Plate Number and Color Recognition (VLPNCR)

VLPNC refers to the automatic extraction from the image data of license plate and information identification. The report includes Arabic numerals and license plate colors.

The license plate acknowledgment system primarily includes four essential aspects: License plate location extraction, background color acknowledgment, number segmentation, and license plate Arabic number acknowledgment. The latest market survey regarding these steps is critical to building up an efficient working system to fulfill the criteria, it is required to study the newest literature of paper and technologies.

Research study on vehicle number plate suggestion or Automatic Number Plate Recognition (ANPR) is mostly done to produce an intro that has high accuracy. Several techniques of picture managing implemented as side detection as well as morphology, link analysis in between things, artificial intelligence, and additionally deep discovering. The system established likewise included professional system to be able to discover the blunder of the number plate acknowledgment in addition to repair it based on the positioning of the personality group in the number plate. In this research study, a K-NN gadget finding out ANPR system established in character acknowledgment.

Research of license plate recognition based on HSV space suggests an approach based on HSV color acknowledgment of license plate and fixes the fundamental issue of license plate acknowledgement. The license plate detection method constructed on color, because the color of the type of car license is limited, and the characters and the cards have apparent color difference. The primary method advantage of these features for license plate recognition is the color edge algorithm, the color distance and similarity algorithm, and so on the distance similarity algorithm. Accordingly, this paper puts forward the different from most other recognition methods of a vehicle authorization plate. The system on the HSV color space is a unique advantage and the color characteristics of the license plate and integration of the binary image mathematical morphology processing method.

Language recognition and translation from document [6] are a crucial action of a document analysis system, considering that recognition engines call for the combination of a language version to boost the transcription performance. The application can scan input message from a file, segment the text, and compute the self-confidence value, after that detect the types of input language and convert the text into the target language. The project is working for four languages, which are Bangla, English, Spanish, and Arabic, that are done by the Tesseract OCR.

2.2. Image Pre-processing

Pre-processing is the initial phase in the electronic image processing, which improves the high quality of the image information for both proper esthetic perception as well as computational handling. Pre-processing improves the image information by getting rid of both background noise, undesirable information, and image reflections as well as normalizing the intensities of the individual image fragments. A significant reason for the failing of vehicle license plate discovery is the poor quality of the vehicle image information. The aim is to remove noise with out losing quality of data needed. Here are some image processing [7] steps, including: The conversion into grayscale, Grayscale to binary image conversion, Resizing the image, and Gaussian Blur Image.

2.3. Plate Area Extraction

Plate area extraction is the second phase of the projects that is process splits in four steps as (1) Morphological operation, (2) extract all contours and matched as characters, (3) group of contour and possible plate detection, and (4) character segmentation.

2.4. Plate Background Color Recognition with HSV Color Space and Threshold Range

HSL and also HSV are alternative depictions of the RGB shade model. HSV color has three components of hue (H), saturation (S), and brightness (V) from that only H associated with color. Hence, when the value of H does not change, and S and V components have little change, they represent the color range fix. H value, when the value of V is not changed, changes in the brightness of V, and the saturation of S, the saturation, and brightness of color gets change. When the equations of $V = 1$ and $S = 1$ have established, the color has the highest purity. Reduce the value of S; color tends to become white. When the value of V is close to zero, the color becomes dark. Therefore, the saturation of S and the brightness of V affect the final color yet. Here, if the equation of $V > T$ and $S > T$ has established, the color is expressed by the H that is the color so the calculation of hue (H), saturation (S), and brightness (V) threshold in the given range of the final output of recognition of color.

2.5. Optical Character Recognition (OCR)

Optical character recognition alters characters included in the image into characters' format [8] commonly, the text of the checked image consists of published letters, transcribed letters, and so forth. The character is checked to evaluate the image, and afterward, the character is gotten. Digitally refining published and transcribed characters, then storing them on a computer system, browsing them, and also

introducing them is an area of computer system vision and also pattern acknowledgment study. OCR is a complex process that involves many steps. The steps involved in OCR are pre-processing feature selection and classification. First, capture the image of the digit categorized in a standard image format such as JPEG, PNG, or bitmap.

2.6. Machine Learning (KNN Algorithm)

Machine learning guided to examine datasets to generalize as well as likewise observe the patterns of that information or details. To anticipate the future worth or behavior from those checking or designs, it will absolutely after that iteratively gains from information, unlike normal computer system programs. The purpose of machine learning is to program computer systems to make use of instance information as an experience or variant as well as additionally use the patterns of these details to anticipate the future based on that information. Machine learning does not just deal with details resource issues, in addition to it is in addition an application of professional system (AI) also. It aids to resolve various troubles in face recommendation, biometrics verification, clinical diagnoses, farming, economics, and robotics [9], [10]. Machine learning involves training a computer version with data or historical info [11], to potentially forecast behavior of the system in the future. Machine learning split right into three primary parts: (1) Supervised learning, (2) unsupervised learning, and (3) reinforcement learning. [Figure 1]

2.6.1. Supervised learning

Supervised learning consists of historic forecasters usage as well as end results with the intent that the model offers valuable predictions of brand-new combinations of forecasters [12] Supervised finding out formulas come in lots of kinds with details toughness, weak points, as well as objectives [11]. Details versions that are suitable for the research study in this paper include linear regression and random forests [13].

2.6.2. Define train-test data

Training machine learning model, the dataset must be separated into two datasets: (1) Training dataset and (2) testing dataset. The separation into two sets is significant since the training process forms the basis of the ability of the procedure to generalize, which is measured by performance on the testing dataset. Models developed to manipulate the training dataset. After that, it makes predictions on the testing dataset. The fundamental design shows in Fig. 2a.

2.6.3. k-Nearest neighbor (kNN) algorithm

The k-nearest neighbor method (kNN) describes a strategy to classify objects in the functions space based on the nearby training samples. KNN is sample-based discovering. Features

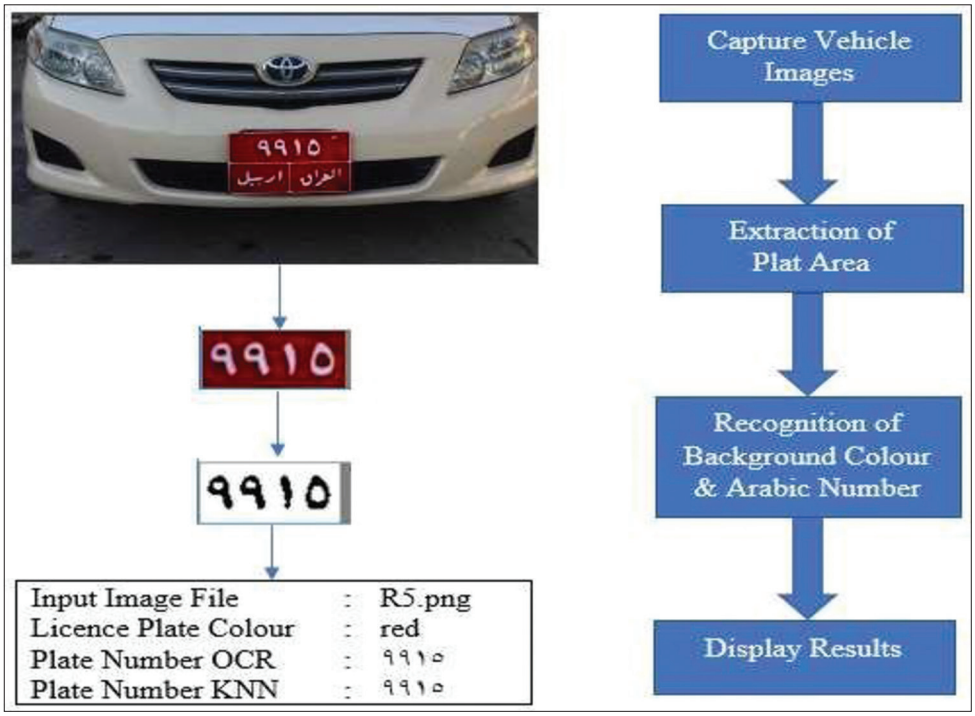


Fig. 1. Fundamental diagram of the research work.

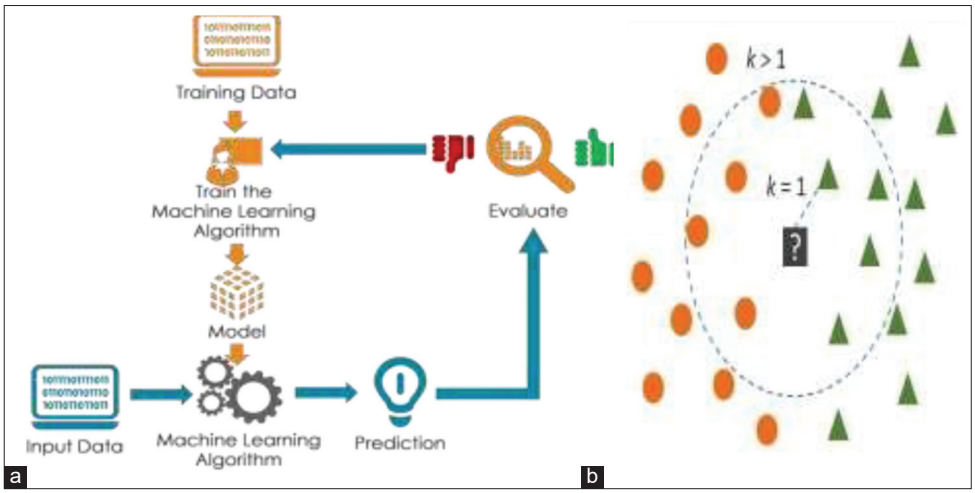


Fig. 2. (a) Training – testing based classifier and (b) kNN Algorithm of distance.

are exclusively approximated in your area in addition to every computation postponed till classifications. KNN procedure thought about the most convenient contrasted to various other machine learning algorithms. KNN as character classifier is a reputable as well as thoroughly used classifier on the market as a result of its simple [5], [14] considering that it does not believe any sort of styles for the circulation of characteristic vectors in space, it additionally has high error resistance over non-linear multi-class issues. KNN computes the range from an unknown to all samples in the design room and also keeps in mind the marginal variety [15]. The class that has the very little array from the unknown is the closest next-door neighbor k= 1, where the value k refers to the neighborhood dimension (Fig. 2b). When it pertains to k > 1, voting of the bulk decision is made to establish the class of the unknown example. It can be seen from [1 that k = 1 continuously creates the highest possible accuracy in addition to is verified by trial and error through the advancement procedure.

3. METHODS AND MATERIALS

This project is a technological remedy for vehicle photos making use of various techniques of image processing. The main target of the system is to give a computerized service utilizing artificial intelligence, image processing, and machine learning for achieving sustainability in the area of transportation or even more generally called as intelligent transportation systems (ITS). However, the system greatly relies on catching these pictures in a top-quality way. The project implemented in Python uses various software and libraries such as OpenCv, Google Tesseract, pytesseract, imutils, Numpy, Scipy, Sk-image, and Sklearn. The project implementation flow chart is shown in Fig. 3.

3.1. Dataset Collection

Project experiment required input vehicle images that contain visible license number plate with proper minimum resolutions. Furthermore, it should contain all type of background color, and Arabic number plates to fulfill the requirements its total of 90 images from those 46 images captured by CCTV and 46 images downloaded through internet. The images contain all types of background color available number of plates. The list of images with background color is given in Table 1.

3.2. Plate Area Extraction

The morphology method is the method most often used to extract a character from the image. The area containing the license plate part stood for as a region of passion (ROP). With the assistance of morphological procedure, finding all possible contours as characters using pre- processing steps. Morphology applied to a binary and grayscale representation. The purpose of the removal is to obtain the characters on the vehicle plate. Each contour location indeed is scanned to get prospects of personality. Each candidate of the personality undoubtedly inspects whether this is a personality or not.

The number plate extracted for further processes such as character segmentation group of similar characters. The algorithm carried out on all contours, which is used to detect the boundaries and edges of the character relevant code. The license plate area number of the edge formed by group of all matched contours edge detection methods. Crop the

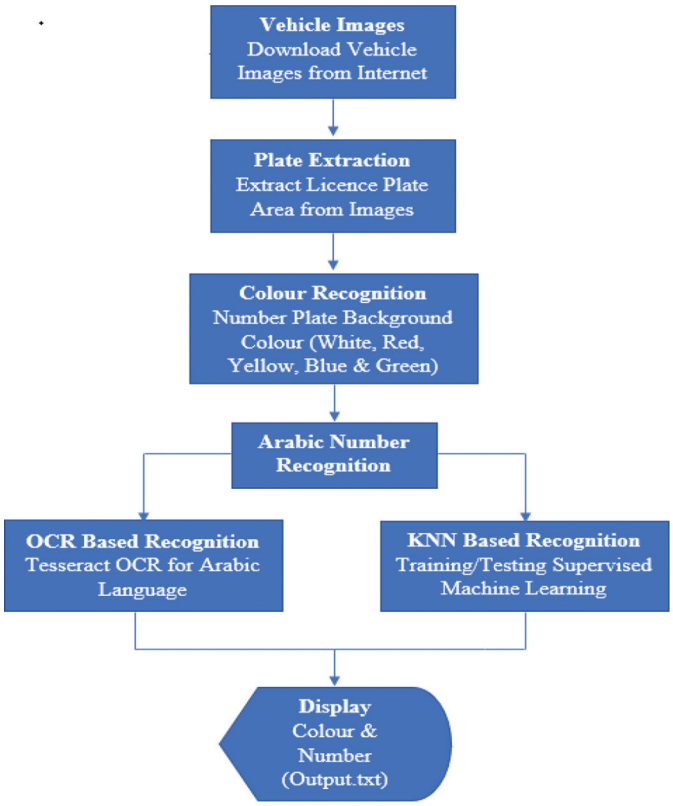


Fig. 3. Research work diagram.

particulate contours group edge from the original image, so its RGB extracted a license plate which contains only Arabic number. Find all possible plates by minimum group and matching contours and make a list of all plates.

The whole steps of implemented as code in a project for Arabic number parts are shown in Fig. 4.

3.3. Implementation of Color Recognition

Once the license plate has been extracted, each plate background color can be recognized. Plate color is characterized by an initial RGB color. The first step is to convert the RGB image into HSV space because the affected light conditions need image histogram equalization. Next step is to calculate the hue (H), saturation (S), and brightness (V) threshold range between lower and upper limit to find the plate color. The resulting range with fuzzy logic is the output of the plate color. The color recognition algorithm implements in "Recognition.py" as shown in Fig. 4 and the color recognition sample, as shown in Table 2.

TABLE 1: License plate background color list

Background color	Vehicle type	No. of Images
White	Private vehicle	36
Yellow	Trucks, tractors, and cranes	18
Red	Vehicle for hire, bus, and taxi	13
Blue	Governmental	09
Green	Army	08
Black	Temporary plates	06
Total images of vehicle plates		90

TABLE 2: Color recognition for all six types of vehicle

Plate						
Color	Black	Blue	Green	Red	Yellow	White

3.4. Implementation of OCR using Pytessrect

Tesseract OCR is precise open-source optical character recognition engines for the Arabic language. Pytesseract is the Python library for Tesseract OCR that can use Python script on Tesseract OCR. The Tesseract OCR library is used to analyses the extracted plate area, and character string is retrieved from characters. Pytesseract image to string function and image to data with specific configuration output is OCR text and confidence level. The Tesseract OCR algorithm implements in "Recognition.py".

3.5. Implementation of KNN Algorithm

The purpose of segmentation is used to extract the target of interest from an image character identity. It is creating the character database taking place on a pre-prepared set of characters containing all necessary for correct identification Arabic numbers; the reference image should be prepared in a format and with parameters such as a scanning system. A set of characters for learning appears on the screen. Here, the table shows the different font size sets of Arabic numbers. The created database of the letter inside the "letter" directory inside the project directory. That contains all ten directories of Arabic numbers (0–9) and samples of images. The image size of each letter is 30 x 45 as dimension where 30 widths and 45 heights of each number image. The number-letter directories and images inside each letter, as shown in Fig. 5.

K-Training samples, whose qualities are reasonably comparable (closest) to the test samples acquire. The examination examples are categorized based on the course labels of the most intimate training samples. These training samples are known as the nearest neighbors. The nearest neighbor (NN) formula called an instance-based approach as

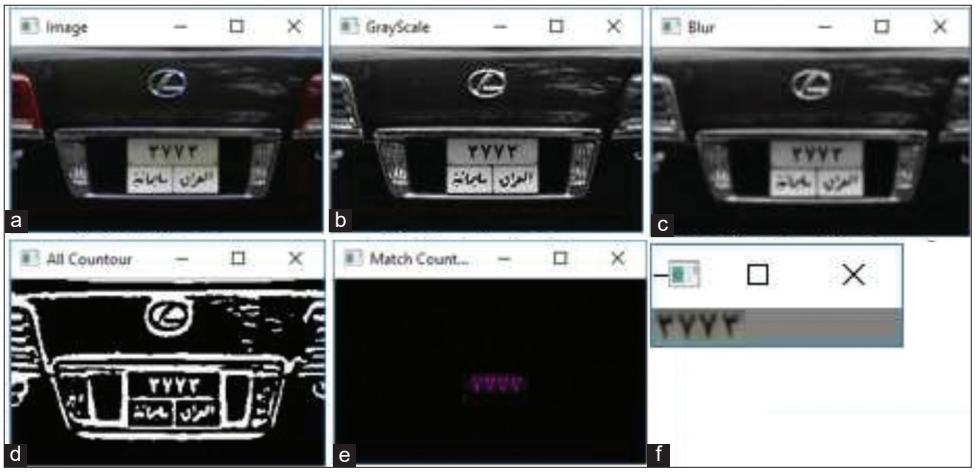


Fig. 4. Whole steps of plate area extraction of the vehicle image. (a) Vehicle image in RGB, (b) Gray-scale, (c) Gaussian blur by 3 × 3, (d) All contours edge by morphology, (e) All matched contours and group of contours, and (f) Extract plate area. More detail explanation in the other answer file.

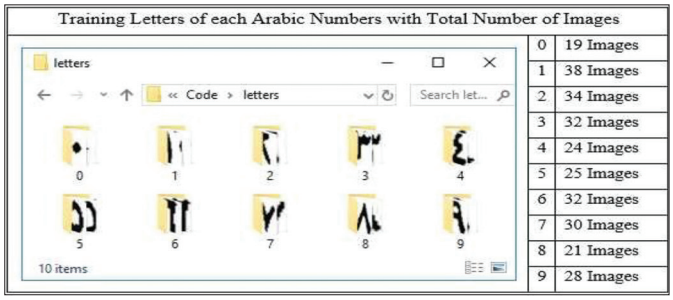


Fig. 5. Arabic number images as a training database.

the examination instances compared to training instances that have actually stored in memory. Here, the training model is created from the 0–9 Arabic numbers, which shows in figure, and the model saved as "knn_model.pkl".

3.6. Writing the Results as Files

The project has successfully extracted plate area and recognizes plate color and Arabic number with confidence by OCR and KNN techniques. The result of recognition displays, as shown in Fig. 6, as the text file (A) "NumberPlate.txt" and (B) on command prompt. Furthermore, written results in the form of the database as "NumberPlate.data."

4. RESULTS AND DISCUSSION

The outcomes of methods that are widely used by researchers to detect the location of the vehicle plate are discussed in this section. Starting with taking vehicle objects and plate locations and recognizing the system. The system checks the area where it undoubtedly the license plate number characters. Each image has tested whether it is a vehicle number plate or not. Furthermore, recognizing plate background color and Arabic number on the plate by KNN and OCR methods.

4.1. Plate Area Extraction

In the beginning, our objective discovers a depictive Arabic number set from number plates, which are identifiable by human beings. The system has tested on all 90 images to extract the plate area. The sample of plate area extraction results as shown in Fig. 7.

4.2. Background Color Recognition

The second goal of the project is to find the background color of the plate to identify vehicle types. Once the vehicle plate area is detected, we can classify it based on its plate colors and detect vehicle type. Intensity correlates with visual image quality, contrast, brightness, etc. Furthermore, taking image time as day time and night time affects the results. The vehicle license plate is mostly a combination of light background and darker characters, otherwise dark background and bright characters. The system implements an HSV mode color image for the detection of plate color instead of RBG.

Conduct the testing of the experimental project on the whole 90 images for plate background color, the results of the recognized color are shown in Table 3. Explained in more detail in the other answer file.

4.3. Arabic Number Recognition

The third goal of the project is to understand the Arabic number in the vehicle plate. Each extracted plate has an

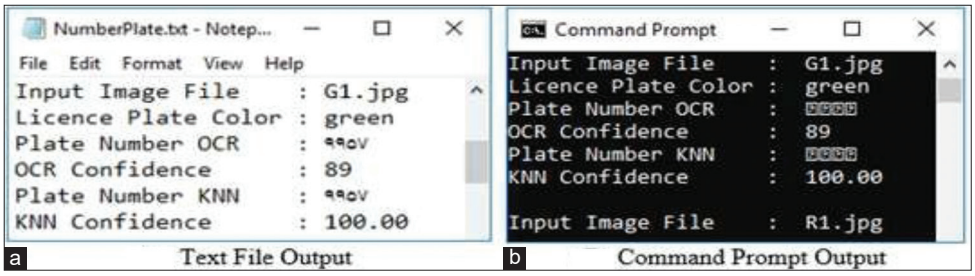


Fig. 6. (a and b) Results as output in text file and command prompt.

Fig. 7. Sample of plate extraction.

Arabic number that is identified using both the technique as KNN and OCR. The results of OCR and KNN for 90 images are shown in Table 3.

4.4. Preliminary Prediction

The project tested on 90 images which have different sizes, dimensions image types, and resolutions. The smallest image size is 7.35 kb, and the highest image size is 804 kb. The minimum pixel dimension of the image as 194 W × 178 H and 219 W × 57 H. Minimum width is 194-pixel and the minimum height is 57 pixels. The maximum pixel dimension of image is 1920 W × 1080 H and 960 W × 1280 H. Maximum width is 1920 pixels and the maximum height is 1280 pixels. Among all these types of images, plate area extraction by removing Arabic characters and other noise is done successfully by retrieving only Arabic numbers. All those extracted plates have different background colors those areas: (1) Black color (six image), (2) blue color (nine images), (3) green color (eight images), (4) red color (13 images), (5) white color (36 images), and (6) yellow color (18 images). The project has identified all the background colors of the extracted plate efficiently with 97.78% accuracy. Table 4 shows the sample result of all images for OCR and kNN.

4.5. The best technique for Arabic Number Recognition

The obtained all plates tested for Arabic number recognition with OCR and KNN techniques shown in Table 4. OCR identified 41 number plates, and KNN identified 83 number plates correctly. The confidence of identification OCR versus KNN for all images is shown in Fig. 8.

Here, the graph X-axis shows image count and the Y-axis demonstrated the confidence of recognition (OCR: Blue line KNN: Red line). The graph clearly shows that the overall confidence level of attention of KNN is higher than OCR among all images. OCR does not identify two images whose confidence level is −1, whereas KNN identified that also. The performance of Arabic number recognition using the KNN technique is better than the OCR technique.

4.6. Research Performance achieved benchmark

The research aimed to achieve license plate recognition for Iraq vehicles containing number plates of different background color types. An efficient methodology for automatic number plate extraction results in extracting the license plate information from 90 images of an Iraqi vehicle successfully that can be used in a real-time environment. This project discusses the five license plate recognition steps, including (1) image pre-processing, (2) license plate location, (3) Background-color recognition, (4) Arabic number segmentation, and (5) Arabic number recognition.

4.6.1. Plate extraction and background color recognition (97.78 % accuracy)

In the image pre-processing stage, the unwanted character gets removed, gray, and binary processing added, which significantly improves the image accuracy of plate extraction. Second, the background color of the number plate is recognized using the HSV color model instead of using the RGB color model. The detection of background color plate area of extracted plate worked with high accuracy of 97.78 % resulted in recognizing Iraqi vehicle type (Army, Government, Police, etc.) successfully.

4.6.2. Arabic number recognition

Arabic is among one of the most renowned languages in the world. The following task to section the number characters of the License Plate is done after getting drawn

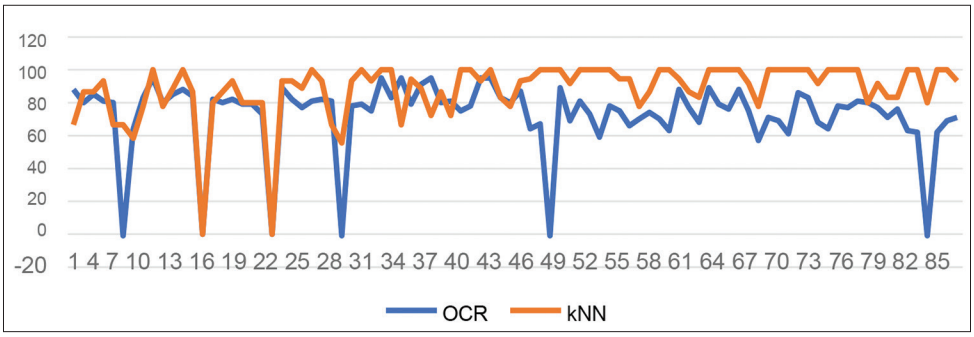


Fig. 8. OCR versus KNN confidence for all the images.

TABLE 3: Plate background color, OCR, and KNN results for all images

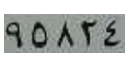
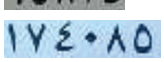
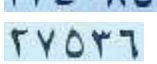
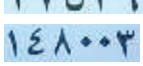
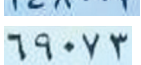
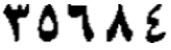
Extract plate	Color	OCR	KNN	Extract plate	Color	OCR	KNN
	black	٠٢٨٢٣	٠٣٢٨٢		Red	٠٧٥١٢	٠٧٥١٢
	Black	٤٨٦٥٣	٤٨٦٥٣		Red	١٣٠٦١٥	١٣٠٦٥١
	Black	٦٦٧١٩	٦٦٧١٩		Red	٨٠٤١٧	٨٠٤٧١
	Black	٣٧٠٥٦	٣٧٠٥٦		Red	٢٧٨١٣	٣١٨٧٢
	Black	٨٤٢٠١	٨٤٢٠١		Red	٨٧١١٣٢	٨٧٢٣١١
	Blue	NA	٣٥٥٦		Yellow	NA	١٧٠٤٨
	Blue	٢١٠	٠١٢		White	٤٢٨٥٩	٤٢٨٥٩
	Blue	٦١٢	٦٢١		White	٥٠٨٤٧١	٥٨٠٤٧١
	Blue	٣٤١	٣٤١		White	٦٣٥٧٢	٦٣٥٧٢
	Blue	٧٩١	٧٩١		White	٣٠٨٠٤١	٣٠٠٨٤١
	Blue	٠٧١	٠٧١		White	٧٣٠٩٦	٣٧٠٩٦

TABLE 4: Sample of OCR and kNN results for all images

No.	Extracted plate	OCR		kNN	
01.		٠٢٨٢٣	✓	٠٣٢٨٢	✓
02.		٤٨٦٥٣	✓	٤٨٦٥٣	✓
03.		٦٦٧١٩	✓	٦٦٧١٩	✓
04.		٣٧٠٥٦	✓	٣٧٠٥٦	✓
05.		٨٤٢٠١	✓	٨٤٢٠١	✓
06.		NA	×	٣٥٥٦	×
07.		٢١٠	×	٠١٢	✓
08.		٦١٢	×	٦٢١	✓
09.		٣٤١	✓	٣٤١	✓
10.		٧٩١	✓	٧٩١	✓

out Arabic License Plate. The method tested on about 90 vehicles obtained plate images of different background colors for Arabic number recognition by OCR and KNN technique. In our experiments, it has observed that among all 90 plate images, OCR technology recognized the Arabic number accurately for 65 images but some images show sequence irregular so finally 41 images identified perfectly with 45.56% accuracy while KNN technology recognized Arabic name for 83 images with 92.22% accuracy. The performance of KNN technology is better than OCR.

4.7. Discussion

By comparison the proposed system with [22] that used machine learning technique to detect and recognize Iraqi plate number, as a result, determined that the accuracy and confidentiality is less than proposed system. Table 5 showing the comparison result details.

TABLE 5: Comparison results

K Parameter	No. of Image	No. of Successful Result	Accuracy
The Proposed System	90	83	92.2%
D. AbdAlhamza, A. Alaythawy [16]	50	45	90%

5. CONCLUSION

This project implements the five steps of license plate extraction, including image pre- processing; license plate location; background color recognition; character segmentation; and character recognition. In the pre-processing stage, GRAY and BINARY processing is added, which significantly improves the image accuracy. Then, the spatial domain filtering is used to get an accurate location for the license plate area. In the license plate positioning stage, the position is determined by edge detection. The license plate is accurately positioned and cropped. The project implements a complete developed system to recognize the background color of the plate using the HSV color model and color histogram depending on finding a matching range of color in between the lower and upper fields. It successfully identifies each six types of background color as white, red, green, yellow, blue, and black. The project proposes an approach for the recognition of Arabic numbers in various scenarios and complex scenes Google Tesseract is specified for character recognition in an OCR technology, generating automated number plate acknowledgment. Supervised machine learning and template matching technique as KNN is one of the most advantageous methods carried out in the proposed project. It helps in recognizing the Arabic number with high accuracy and makes the identification closer.

These proposed project experimental results show that accuracy for each module is achieved up to the greatest extent on the dataset of the total 90 images. The project has achieved a plate extraction and background color accuracy as 97.78%, Arabic number recognition with OCR as 45.56%, and with KNN as 92.22%. On comparison of various Arabic number recognition algorithms, it inferred that the KNN technique is the most efficient one. Thus, the proposed system achieves the extraction of number plates and the recognition of Arabic numbers in vehicles. It works out better using a low-cost embedded board and software such as python. The project demonstrates that image processing is a far more efficient method of license plate recognition as compared to traditional techniques.

REFERENCES

[1] I. N. Mahmood, H. T. S. AlRikabi, A. H. M. Alaidi and F. T. Abed. "*Design and Implementation of a Smart Traffic Light Management System Controlled*" Wirelessly By Arduino, 2019.

[2] Y. Kessentini, M. D. Besbes, S. Ammar and A. Chabbouh. "A two-stage deep neural network for multi-norm license plate detection and recognition". *Expert Systems with Applications*, vol. 136, pp. 159-170, 2019.

[3] S. Sugeng and E. Y. Syamsuddin. "Designing automatic number plate recognition (ANPR) systems based on K-NN machine learning on the raspberry pi Embedded system". *JTEV Journal Teknik Elektro dan Vokasional*. vol. 5, pp. 19-26, 2019.

[4] Q. Ying and J. G. Sheng. "Research of License Plate Recognition Based on HSV Space". *3rd International Conference on Materials Engineering, Manufacturing Technology and Control*. Atlantis Press, 2016.

[5] A. R. Quiros, R. A. Bedruz, A. C. Uy, A. Abad, A. Bandala, E. Dadios and A. Fernando. "A kNN-based Approach for the Machine Vision of Character Recognition of License Plate Numbers". *TENCON-IEEE Region 10 Conference*, pp. 1081-1089, 2017.

[6] M. Khan, M. Hassan, A. B. M. Noman and S. Rajbangshi. "*Language Recognition and Translation from Document*", 2018.

[7] A. Al-Zawqari, O. Hommos, A. Al-Qahtani, A. A. Farhat, F. Bensaali, X. Zhai and A. Amira. "HD number plate localization and character segmentation on the Zynq heterogeneous SoC". *Journal of Real-Time Image Processing*, vol. 16, pp. 1-15, 2018.

[8] E. Alpaydin. "*Introduction to Machine Learning*". MIT Press, Cambridge, 2014.

[9] M. Mohri, A. Rostamizadehand and A. Talwalkar. "*Foundations of Machine Learning*". MIT Press, Cambridge, 2018.

[10] P. Lison. "*An Introduction to Machine Learning*". Springer, Cham, Denmark, 2015.

[11] M. Kuhn and K Johnson. *Applied Predictive Modelling*". Vol. 26, Springer, New York, 2013.

[12] T. Hastie, R. Tibshiraniand and J. Friedman. "*The Elements of Statistical Learning*". Springer Series in Statistics, Springer, New York, 2001.

[13] J. O. Rawlings, S. G. Pantula and D. A. Dickey. *Applied Regression Analysis: A Research Tool*. Springer Science and Business Media, Germany, 2001.

[14] T. K. Hazra, D. P. Singh, and N. Daga. Optical Character Recognition using KNN on Custom Image Dataset. In: *8th Annual Industrial Automation and Electromechanical Engineering Conference*, pp. 110-114, 2017

[15] N. B. A. Hamid and N. N. B. Sjarif. "Handwritten Recognition Using SVM, KNN and Neural Network". *arXiv*, vol. 2017, p. 00723.

[16] D. AbdAlhamza and A. Alyathawy. "Iraqi license plate number recognition based on machine learning". *Iraqi Journal of Information and Communications Technology*, vol. 3, pp. 2222-758x, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Molecular detection of Enterotoxigenic *Escherichia coli* Toxins and Colonization Factors from Diarrheic Children in Pediatric Teaching Hospital, Sulaymaniyah, Iraq

Hezhan Faeq Rasul, Sirwan Muhsin Muhammed, Huner Hiwa Arif, Paywast Jamal Jalal

Department of Biology, College of Science, University of Sulaymaniyah, Sulaymaniyah, Iraq

ABSTRACT

Enterotoxigenic *Escherichia coli* (ETEC) is one well-established causative agent of diarrhea in the developing countries among young children. This prospective study was performed at Laboratories of University of Sulaimani (in Sulaymaniyah City/Iraq) from September to October 2021which aimed to determine the prevalence of ETEC among children and the most prevalence colonization factor (CFA/I) among ETEC. One hundred and twenty-five fresh stool samples were collected from hospitalized – children with diarrhea at Dr. Jamal Ahmed Rashid's Pediatric Teaching Hospital. The collected samples were cultured on MacConkey and eosin methylene blue agar as selective and differential media for Gram- negative bacteria. Colonies were identified through Gram staining and biochemical tests including: Indole, methyl red, and catalase reaction test. Vitek-2 machine was depended to test some obtained isolates. Most of isolates (60%) showed positive results for *E. coli* – out of this percentage, 14 (18.66%) were positive for ETEC using polymerase chain reaction assay identifying stable and labile toxins (LTs). It was noticed that all of the ETEC isolates were stable toxin producer isolates whereas LT producer isolates were not identified. Colonization factor 5 (CS5) has been detected among three ETEC isolates (21.42%), meanwhile, 11 isolates (78.57%) have not expressed colonization factors at all.

Index Terms: *Escherichia coli*, Enterotoxigenic *E. coli*, Stable toxin, Labile toxin

1. INTRODUCTION

Bacterium coli commune was initially reported as a commensal Gram-negative rod from the healthy individual's intestinal flora by Theodor Escherich, a German pediatrician, in 1885, and in his honor, these rods were named *Escherichia coli* [1]. The genus *Escherichia coli* is distributed widely and is the most common facultative anaerobe found among humans and warm-blooded animals' – large intestine [2]. Depending on the number of virulence determinants found, specific combinations were created, determining the currently known *E. coli* pathotypes, which are generally recognized as diarrheagenic *E. coli* (DEC) [3]. DEC pathotypes are classified into enteropathogenic *E. coli*, enterotoxigenic *E. coli* (ETEC), EIEC is for enteroinvasive *E. coli*, and Shiga stands for enterohemorrhagic *E. coli*, enteroaggregative *E. coli* (toxin-producing *E. coli*) is another type of *E. coli*. They pathotypes of *E. coli* vary widely in terms of preferred host colonization locations, virulence mechanisms, and clinical symptoms and out [4]-[6].

In the developing countries, ETEC is still one of the most common causes of infectious diarrhea in travelers and children [7]. Watery diarrhea, vomiting, stomach cramps, and,

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp49-57

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Hezhan Faeq Rasul, Department of Biology, College of Science, University of Sulaymaniyah, Sulaymaniyah, Iraq. E-mail: hezhan.rasul@univsul.edu.iq

Received: 16-06-2022 Accepted: 31-08-2022 Published: 22-09-2022

in some circumstances, declining in body temperature are the common symptoms of ETEC infections [8]. Infections can be self-limiting in normally healthy people, but they may be fatal among children and young adults as well as among immune compromised patient [9]. ETEC causes over 200 million cases of diarrhea and 380,000 deaths per year, mostly below the age 5 among children [10].

ETECs ability to stick to and colonize intestinal epithelium, it is critical for pathogenicity. In addition, its ability to produce heat-labile labile toxin (LT) enterotoxin and/or heat-stable stable toxin (ST) enterotoxin, both of which can produce diarrhea. ST is a limited peptide made up of 18–19 amino acid residues, whereas LT is a high-molecular-weight (84 kDa) enterotoxin with an active alpha subunit surrounded by five identical binding B subunits [11]. The two main genotypes of ST are STa and STb; typically, ETEC strains isolated from people produce STa (STI or STh), which is encoded by the *estA* gene, whereas STb (STII orSTP) is primarily produced by animal ETEC strains which is encoded by the *estB* [12]. The LTs that ETEC strains produce are likewise a diverse category of toxins. There are two main LT families known as LT-I and LT-II [13]. LT genes *eltA* and *eltB* produce LT-I and LT-II, respectively. The ST genes are possible to express independently or in tandem with the LT genes *eltA* and *eltB* [13]. ETEC strains can express seven different toxin combinations: STh, STp, STh/LT, STp/LT, LT, and less typically, STh/STp and STh/STp [14]. The existence of colonization factors (CFs) on membrane of a bacterial cell, which normally form pili, also known as fimbriae, is necessary for colonization [15]. Depends on antigenic specificity and/or the N-terminal amino-acid sequence of the main subunit, different forms of colonization factor antigens (CFA) and putative colonization factors have been identified(pilin) such as CFA/I, CS1, CS2, CS3, CS4, CS5, CS6, CS7, CS14, CS17, CF19, CF21, CF22 [16]. There is a limited research available on ETEC colonization and prevalence of diarrheagenic ETEC among human, particularly children below the age of six in this region. To fill this gap, this study have conducted to determine *E. coli*, ETEC toxins (ST and LT) producers, and colonization factors from children under 6 years suffering from watery diarrhea in the Pediatric Teaching Hospital, in Sulaimani City.

2. MATERIALS AND METHODS

2.1. Sample Collection

During the period from September to October, 125 sample stools were collected from children <6 years at Dr. Jamal Ahmed Rashid's Pediatric Teaching Hospital and SMART Private Hospital in Sulaymaniyah City. Both sexed were included (63 females and 62 males). The necessary information about the patients were taken from the hospitals, and the collected samples were transferred from the hospitals to the Advanced Bacteriology Laboratory from Biology Department of University Sulaimani within less than 3 hours in an ice box to culture them.

2.2. Bacterial Cultivation and Characterization

All samples were preliminary cultured on differential and selective media for presumptive isolation of Gram- negative enteric rods. These included MacConkey agar for first isolation and eosin methylene blue for confirmation as described by [17], [18]. All lactose fermenting, deeply pink, circular, medium in size, colonies were subcultured on the medium (eosin methylene blue) agar (Neogene, UK). All plates were incubated at 37°C for 18–20 h, Colonies showing green metallic sheen on EMB agar were considered as *E. coli* strains. *E. coli* samples utilized in the present study were identified by Gram staining [19] and initial biochemical tests including indole [20], methyl red [21], catalase test [22], and other bacteriologic characterization using Vitek-2 system (VITEK®2 GN ID card) by Vitek machine (BioMerieux, France) were performed for some of them [23].

2.3. DNA Extraction and Purification

The DNA of isolates under test was isolated and purified using and following the directions.

2.3.1. Colony extraction

It was performed by transferring two colonies from fresh bacterial culture then mixed with 40 µl of ddH_2O and preheated at 95°C for 10 min using the thermo cycler and purified DNA obtained by centrifugation at 12,000 rpm for 1 min. The supernatant was used as a polymerase chain reaction (PCR) template [24].

2.3.2. DNA extraction with kit

Overnight fresh colonies from nutrient broth (Neogene, UK) utilized. Genomic DNA from *E. coli* isolates was extracted and purified using a DNeasy kit (AddPrep Genomic, Korea) according to manufacturer protocol.

2.4. PCR Method

PCR mixture contained the DNA template, forward/reverse primer (Macrogen, Korea), and Master Mix (*Taq* Master (2 × conc.)/addbio. Korea) deionizing water (Accumax, Korea).

2.4.1. 16S rRNA

PCR was performed for 75 samples *E. coli* to identify 16S rRNA using this reaction included: Initial denaturation for 5 min at 94°C, followed by 35 cycles of amplification (1 min at 94°C, 1 min at 56–58°C and 1 min at 72°C), and finally finished with 7 min at 72°C [25].

2.4.2. ST and LT

The 96-well plates were used to amplify stable and LT genes. The PCR procedure included pre-incubation at 95°C for 1 min, followed by 35 cycles of (1 min at 95°C, 1.10 min at 45°C, 1.30 min at 72°C), final incubation at 72°C for 5 min [26]. The products have run on 2% agarose gel (TransGen, China).

2.4.3. Colonization genes

To identify genes of colonization factors of CFA/I, CS1, CS2, CS3, CS4v, CS5, CS6, CS14, CS17. The same pre-mentioned procedure was depended with different primer for each gene. The genes were amplified by an initial denaturation at 94°C (1 min), followed by 35 cycles of amplification (94°C for 30 s, 52°C for 30 s, and 72°C for 1 min), finally, 5 min at 72° [27]. The amplicon was separated with 3% agarose gels by gel electrophoresis (Cleaver-CS-300v, UK) and then visualized by ethidium bromide (TransGen, China). The specificity of the primers was tested by both BLAST search and is illustrated in Table 1.

2.5. DNA Sequencing

The sequencing was performed for 10 samples with amplified 16S rRNA -F and 16S rRNA-R (forward and reversed primers (10 pmol). DNA sequencing was achieved by Sanger sequencing/ABI 3500, Macrogen Genome Center, Korea using BigDye kit.

2.6. Phylogenetic Tree

Evolutionary analysis was conducted by MEGA7 program. The evolutionary history was deduced using the Kimura 2-parameter model and the maximum likelihood technique [28]. It is shown the tree with the greatest log likelihood (−505.1852). Next to the branch is the proportion of trees where the related taxa clustered together. The starting tree(s) for the heuristic search were automatically generated by applying the neighbor-join and BioNJ algorithms to a matrix of pairwise distances calculated using the maximum composite likelihood technique and thereafter picking the topology with the best log likelihood value. To represent evolutionary rate differences across sites (5 categories [+G, parameter = 0.0500]), a distinct gamma distribution was utilized. The branch distances are calculated by the number of replacement per location, as well as the tree is depicted to scale. A total of 18 nucleotide sequences were examined. The codon locations were included 1^{st} + 2^{nd} + 3^{rd} + non-coding. All positions containing gaps and missing data were eliminated. There were a total of 320 positions in the final dataset [29].

3. RESULTS AND DISCUSSION

3.1. Detection of DEC

It was appeared that out of 125 tested samples, 83 (66.4%) were Gram-negative bacteria after they were cultured on

TABLE 1: Reference strain, primer sequence, number base pair of 16S rRNA, ST, LT toxin and colonization factors

Primer name	bp	Primer sequence	Reference
16SrRNA 16S-F 16S-R	426	5′GACGTACTCGCAGAATAAGC-3′ 5′-TTAGTCTTGCGACCGTACTC-3′	[25]
St toxin STF STR	186	5-TCT GTA TTG TCT TTT TCA CC-3, 5-TTAATA GCA CCC GGT ACA AGC-3,	[26]
Lt toxin LTF LTR	273	5-ACGGCGTTACTATCCTCTC -3 5-TGGTCTCGGTCAGATATGTG -3	[27]
CFA/I	170	5-GCTTATTCTCCCGCATCAAA-3 5-ACTTGTCCTCCCCATGACAC-3	[27]
CS1	243	TCCGTTCGGCTAAGTCAGTT CCGCACATTTCCTGTGTTCT	[27]
CS3	100	CTAGCTTTGCCACCACCATT GGCAACTGACTCCCATTTGT	[27]
CS5	226	TCCGCTCCCGTTACTCAG GAAAAGCGTTCACACTGTTTATATT	[27]
CS4	198	ACCTGCGGCAAGTCGTTT TCTGCAGGTTCAAAAGTCACA	[27]
CS6	152	CTGTGAATCCAGTTTCGGGT CAGGAACTTCCGGAGTGGTA	[27]
CS14	162	TTTGCAACCGACATCTACCA CCGGATGTAGTTGCTCCAAT	[27]
CS17	130	GGAGACGCTGAATACAACTGA CTCAGGCGCAGTTCCTTGTCS2	[27]
CS2	368	AGTGGTGGCAGCGAAACTAT TTCCTCTGTGGGTTCTCAGG	[27]

EMB and MacConkey agar, they showed metallic shine and pink color respectively. Seventy-five (60%) of them were rod shape purple color, when they were grown in peptone water, they produced forming pink ring color at top of tubes after addition of Kovac's reagent. The color of the broth cultured changed to red after adding methyl red indicator to tube during performing methyl red test. H_2O_2 was added to fresh colonies, bubble formation indicated positive catalase test. The percentage of appeared *E. coli* similarity to an ideal *E. coli* by Vitek-2 test was done for the samples of 70, 23, 60, and 35 which were 99%, 93%, 87%, and 94%, respectively. Seventy-five isolates have given positive for 16S rRNA- based PCR, as shown in Fig. 1.

Our study explored that the most diarrheagenic pathogens among Gram negative in Sulaimani are *E. coli*, which is compatible with the results reported in a local study by Hasan *et al*. (2020) done in Dhok city [30], Shatub *et al*. (2021) found similar results (61.3%) [31], whereas Khalil (2015) in Baghdad reported lower positive rates (38.6%) [32] as well as other investigators who showed lower positive results [33]-[35].

Several studies from worldwide revealed varying DEC detection rates in *E. coli* among children under 5 years old, ranging between 4% and 87% in Africa including 22.9%, 7.4%, 55.9%, and 86.5%, Asia (45.2%, 4.7%, 6.82%), and America 5.5%. These variations could be related to changes in DEC pathotype distribution from one region to the others, also between countries in the same region [36]. According to many reports around the world, various factors may be the primary causes of diarrheal outbreaks including; traveling to tropical zones, consuming contaminate, and lack of personal hygiene [35]. However, considering that prior studies have focused on certain aspects such as geographical conditions, sampling period, study population, hygienic level of region, and detection technologies [37].

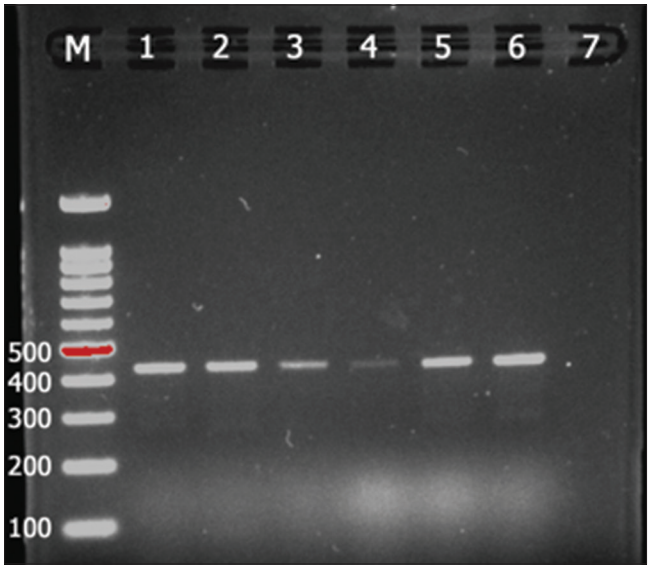


Fig. 1. 16S rRNA gene PCR of *Escherichia coli*: M; is 100 bp ladder 2–6 were 16S rRNA gene.

The proportion of infected males 44 (58.6%) was relatively higher than females 31 (41.3%), the infected males higher than females were like to result reported by Hasan *et al*. (2020) who reported 87.4% among males and 87.0 among females [30]. The current observations were agreed with results reported by the result mentioned by Amir *et al*. (2020) in Iran who showed 53.01% for male and 46.99% for female [35]. Similarly, our observations were parallel to the results concluded by Ochien and Atieno (2021) in West Kenya (55.9% and 44.1%) male and female, respectively [38], whereas the current results were not agreed with the result of Abbasi in Iran (2020) who reported higher rates among females than males [34].

3.1.1. DNA sequencing

All accession numbers have shown in Fig. 2. AY342058.1 is the accession number of a ST gene which sequencing was performed.

A phylogenetic tree based on the 16S ribosomal RNA sequences was extracted from a representative set of 10 *Enterobacteriaceae* genomes and compared to some other different strains (Fig. 2). All strains referred to one clad, the clad of *E. coli* was more similar to *Shigella flexneri*. The numbers (close individual) clustered and bootstrap percentage of 100 replications. Some of the tree nodes are uncertainly predicted. It has concluded that the analysis of variable genes identifies interstrain relationships that may be correlated to the lifestyle of the organisms [39].

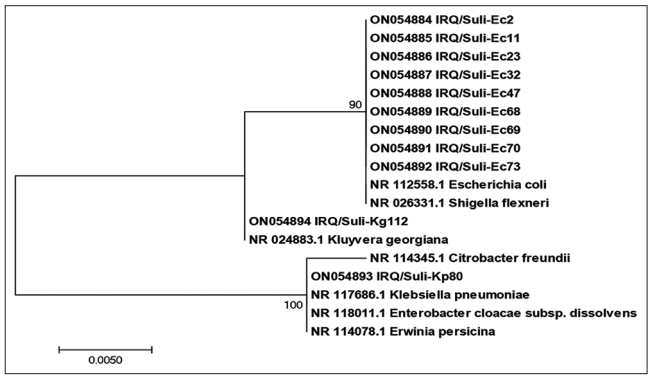


Fig. 2. Molecular phylogenetic analysis by maximum likelihood method.

3.2. Identification of ETEC Toxins

The ETEC characterization using PCR among isolates was done. For the tested children who suffered from non-bloody, acute diarrhea due to *E. coli* (n = 75). All *E. coli* isolates were evaluated by PCR monoplex for ST gene and LT gene. A total of 14 (18.66%) were proved ETEC. The majority toxin profile among the selected strains were ST, whereas LT was not recorded in the present study, as shown in Fig. 3. While all other patho type of *E. coli* were 81.33%.

Results of the present study were agreed with observations reported in studies done in other parts of Kurdistan and Iraq. PCR-based studies detected showed different percentage rates of ETEC in stool samples ranging from 18% to 26% [30], [32], [36], [40], [41]. Results of this study were not parallel to conclusions mentioned by other investigators who reported lower percentage rates of positive results [34], [37].

ETEC is more common among low- and middle-income individual states, where it is a prominent pathogenic strain in travelers' diarrhea, with a large burden on these countries [42]. ETEC was recognized as a major pathotype among children below 5 years old. The high percentage rates of ETEC- positive results could be attributed to the family's poor hygiene and artificial feeding [30], [41], [43]. Variations in our results with other researches could be related to changes in primers utilized, geographical considerations, population targeted, and sample size [44]. In nine of the 12 research conducted in Africa, 22 of 34 investigations in Asia, and three of six studies in Latin America and the Caribbean, ETEC was being the first or second most often isolated pathogen [45].

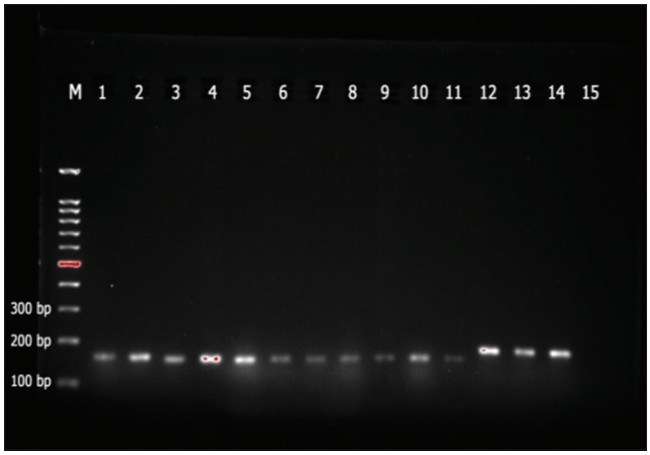


Fig. 3. Agarose gel electrophoresis for ST gene PCR products of outpatient isolates of ETEC. Lane M: 100 bp DNA ladder followed 1-14 are ST Gene, 15: Negative control is distilled water.

The proportion of infected male and female with ETEC was different, in our study, 9 (64.3%) ETEC was among males, while the rest 5 (35.7%) were among females which is parallel to results mentioned by other studies from West Kenya [38] whereas other investigators [46] found higher percentage positive results among females which are not in agreement with our results.

3.3. Distribution of ETEC among Children According to the Age

The ETEC pathotype was identified in all children according to their ages, with a slightly increasing number of infected children with ETEC under 12 months [47]. Our result highlighted the significant of ETEC like a cause of childhood diarrhea between the ages of 1 and 12 months, as shown in Fig. 4. The current observations were agreed to results reported by Shatub *et al.* (2021) from Tikrit/Iraq [31] whereas our results were far with outcome of Khalil (2015) [32]. Our findings are backed up by a review article that looked at ETEC infection from 1984 to 2005, stratified infants by age, and found that the peak incidence occurs after 6 months and can last until 18 months [48]. This might be related to duration of breastfeeding, the source of drinking water, cleanliness, sanitation, age, and the level of maternal educational. In contrast, in the finding by Abdul-Hussein *et al.* (2018) in Wasit/Iraq, the most ETEC prevalence was found among (3–24 months [33].

3.4. Stable and LT

In our result, it was noticed that 14 (18.66%) ST genes were present among isolates while there had no any identified LT solely and ST-LT toxin as shown in Fig. 3. The prevalence of ST was the most common toxin gene. In the other region of Kurdistan and the rest of the world, there are several reports showing differences in the prevalence of ETEC pathotype. The present study differed from a study in Duhok city by

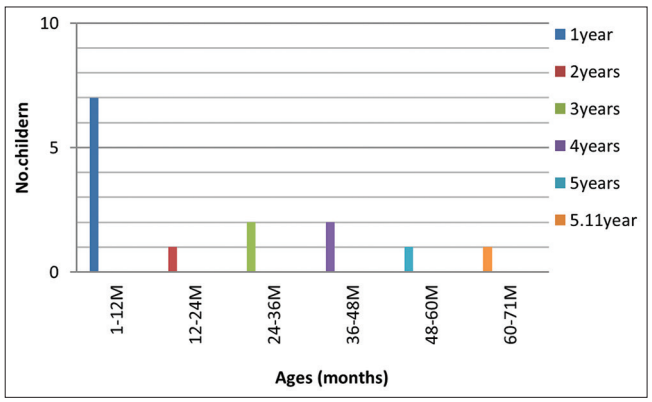


Fig. 4. Distribution of ETEC among children according to their age.

Hasan *et al.* (2020), by which LT toxin solely identified in 37% of the cases [30], whereas seven strains positive for LT gene and three strains positive for ST were identified by Khalil (2015) in Bagdad [32]. Consequently, our study was close to a study by Shabazi *et al.* (2021) that found three ST and one LT gene in their results. Besides, the prevalence of LT and ST gene in several studies was presented as follows. For instance, a study by Alizade *et al.* (2014) found 11.97% ST and 9.86% LT, Saka *et al.* (2019) concluded with 19 ST and 14 LT, and Nazarian *et al.* (2014) with 4 ST and 1 LT [36], [37], [46], [49].

ST, a peptide with a molecular weight of 18 or 19 amino acids, combined with a carrier protein and then will be antigenic. As a result, after infected with ST-producing ETEC, immunological responses to ST are not produced. The percentage of strains that produce LT alone, ST alone, or LT/ST varies by geographic region; in general, 30–50% of clinical ETEC isolates appear to produce ST solely [9], [45].

For pathogenic strains, such as ETEC, it was shown that specific conditions of host might increase or decrease bacterial virulence. The impact of glucose and bile on the gene expression and protein level of ST generated by different ETEC isolates was investigated by Joffre *et al.* (2016), and he discovered that there are unique STa amino acid variations that respond differently to environmental signals such as bile [50].

A substantial amount of literature highlights the effect of different seasons on the prevalence of ETES-associated infections, by which in the late spring and entire summer, this type of infection was repeatedly identified [9], [51]-[54]. However, in our study, the incidence of ETES-associated infection is lower than the expected rate when compared to other studies. This may be due to the period of data collection that was performed in September and October.

3.5. Detection of Colonization Factors

Among the 14 ETEC produced ST isolates, nine primers were chosen for most common CF, among them, only 3 (21.42%) ETEC isolates shown CF, 11 (78.57%) ETEC isolates without CF. In our result, there was only 1 isolate (7%) posed CFA/I, 1 (7%) showed CS4, and also 1 (7%) with CS5 as shown in Fig. 5.

In among clinically important ETEC strains, over 22 antigenically different CFs were identified, but only a handful are frequently present in diarrheic patient samples [9]. Current results are compatible with the result reported by Peruski Jr *et al.* (1999), among the ST- positive strain CFA/1, CS4, and

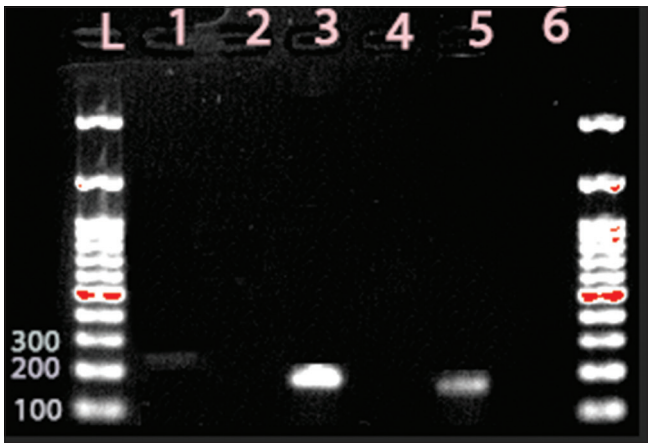


Fig. 5. Agarose gel electrophoresis for ETEC – colonization factor gene identification first lain (M): 100 bp DNA ladder, 1; CS5 which shows 226 bp, 3; CS4 that is 198 bp 5; CFA/1 show 170 bp, 2, 4, and 6 demonstrate negative control contained distilled water.

CS5 were reported, 77% isolates were failed to express any of 9 CF [55]. Our result is close to the finding of Shaheen *et al.* (2004) CFA/I (9.7%), CS4 (2 Strain), and CS5 (2 strain) Shaheen *et al.* (2004) [56] and Kipkirui *et al.* (2021) [44].

The discrepancy between our findings and those of other studies could be attributable of variation in CF expression by ETEC in different geographical regions, as well as differences in laboratory methods/primers used to identify CFs [46]. In addition, due to sub cultured repeatedly or storage for long term, the plasmid containing the CF genes has been lost [44]. Decreasing the expression of CF genes, a mutation inside the genetic locus, and expression of a CF not covered by the primers used in the PCR panel [6], [57], [58]. Antigens for CF are only created *in vivo*, or a small percentage of strains do not generate CFs [59]. The CF antigens are differed from the 9 CFs screened for in this study. Decreasing of CF has been reported to be associated to LT strains [41], [60], which similar our results where lack LT toxin. However, some studies have reported that CFs are almost equally associated with LT- and ST-positive ETEC strains [61].

4. CONCLUSION

This study illustrated ETEC in children below 6 years old with acute non-bloody diarrhea. PCR-based detection of ETEC revealed that all isolated ETECs found with ST toxin-producing gene with no any ETEC -LT identified gene isolate. From the overall of ETEC isolates, only three of them showed CF which are CS4, CS5, and CFA/1 on the different strains. This finding can provide an evidence on the

prevalent of ETECs pathotype in this region, furthermore, creation a platform for vaccine development can be adapted from this finding.

ACKNOWLEDGMENT

We thank to the staff of Dr. Jmale Ahmaed Rashid's Pediatric Teaching Hospital for helping us and we thank to all persons contributed in this study.

REFERENCES

[1] A. D. Mare, C.N. Ciurea, A. Man, B. Tudor, V. Moldovan, L. Decean, *et al*. Enteropathogenic *Escherichia coli*-A summary of the literature. *Gastroenterology Insights*, vol. 12, pp. 28-40, 2021.

[2] S. Ramos, V. Silva, M. L. E. Dapkevicius, M. Caniça, M. T. Tejedor-Junco, Igrejas G, *et al*., *Escherichia coli* as commensal and pathogenic bacteria among food-producing animals: Health implications of extended spectrum β-lactamase (ESBL) production. *Animals* (*Basel*), vol. 10, p. 2239, 2020.

[3] J. P. Nataro and J. B. Kaper. Diarrheagenic *Escherichia coli*. *Clinical Microbiology Reviews*, vol. 11, pp. 142-201, 1998.

[4] G. D. Christensen, W A Simpson, J. J. Younger, L. M Baddour, F. F Barrett, D. M. Melton, *et al*. Adherence of coagulase-negative staphylococci to plastic tissue culture plates: A quantitative model for the adherence of staphylococci to medical devices. *Journal of Clinical Microbiology*, vol. 22, pp. 996-1006, 1985.

[5] Y. Zhou, X. Zhu, H. Hou, Y. Lu, L. Yu, L. Mao, *et al*. Characteristics of diarrheagenic *Escherichia coli* among children under 5 years of age with acute diarrhea: A hospital based study. *BMC Infectious Diseases*, vol. 18, p. 63, 2018.

[6] S. M. Turner, A. Scott-Tucker, L. M. Cooper and I. R. Henderson. Weapons of mass destruction: Virulence factors of the global killer enterotoxigenic *Escherichia coli*. *FEMS Microbiology Letters*, vol. 263, pp. 10-20, 2006.

[7] J. M. Fleckenstein, K. Roy, J. F. Fischer and M. J. I. Burkitt. Identification of a two-partner secretion locus of enterotoxigenic *Escherichia coli*. *Infection and Immunity*, vol. 74, pp. 2245-2258, 2006.

[8] C. K. Porter, M. S. Riddle, D.R. Tribble, A. Louis Bougeois, R. McKenzie, S. D. Isidean, *et al*., A systematic review of experimental infections with enterotoxigenic Escherichia coli (ETEC). *Vaccine*, vol. 29, pp. 5869-5885, 2011.

[9] F. Qadri, A. M. Svennerholm, A. S. G. Faruque and R. B. Sack. Enterotoxigenic *Escherichia coli* in developing countries: Epidemiology, microbiology, clinical features, treatment, and prevention. *Clinical Microbiology Reviews*, vol. 18, pp. 465-483, 2005.

[10] T. P. Madhavan and H. Sakellaris. Colonization factors of enterotoxigenic *Escherichia coli*. *Advances in Applied Microbiology*, vol. 90, pp. 155-97, 2015.

[11] Y. Zhang, P. Tan, Y. Zhao and X. Ma. Enterotoxigenic *Escherichia coli*: Intestinal pathogenesis mechanisms and colonization resistance by gut microbiota. *Gut Microbiota*, vol. 14, p. 2055943, 2022.

[12] I. Bölin, G. Wiklund, F. Qadri, O. Torres, A. L. Bourgeois, S. Savarino, *et al*. Enterotoxigenic *Escherichia coli* with STh and STp genotypes is associated with diarrhea both in children in areas of endemicity and in travelers. *Journal of Clinical Microbiology*, vol. 44, pp. 3872-3877, 2006.

[13] M. A. Lasaro, J. F. Rodrigues, C. Mathias-Santos, B. E. C. Guth, A. Balan, M. E. Sbrogio-Almeida, *et al*. Genetic diversity of heat-labile toxin expressed by enterotoxigenic *Escherichia coli* strains isolated from humans. *Journal of Bacteriology*, vol. 190, pp. 2400-10, Apr 2008.

[14] A. Sjoling, G. Wiklund, S. Savarino, D. Cohen and A. M. Svennerholm. Comparative analyses of phenotypic and genotypic methods for detection of enterotoxigenic *Escherichia coli* toxins and colonization factors. *Journal of Clinical Microbiology*, vol. 45, pp. 3295-3301, 2007.

[15] C. M. Müller, A. Aberg, J. Straseviçiene, L. Emody, B. E. Uhlin and C. Balsalobre. Type 1 fimbriae, a colonization factor of uropathogenic *Escherichia coli*, are controlled by the metabolic sensor CRP-cAMP. *PLoS Pathogens*, vol. 5, p. e1000303, 2009.

[16] O. Puiprom, S. Chantaroj, W. Gangnonngiw, K. Okada, T. Honda, T. Taniguchi, *et al*. Identification of colonization factors of enterotoxigenic *Escherichia coli* with PCR-based technique. *Epidemiology and Infection*, vol. 138, pp. 519-524, 2010.

[17] T. Dadheech, R. Vyas and V. Rastogi. Prevalence, bacteriology, pathogenesis and isolation of *E. coli* in sick layer chickens in Ajmer region of Rajasthan, India. *International Journal of Current Microbiology and Applied Sciences*, vol. 5, pp. 129-136, 2016.

[18] G. Y. Lee, H. I. Jang, I. G. Hwang and M. S. Rhee. Prevalence and classification of pathogenic *Escherichia coli* isolated from fresh beef, poultry, and pork in Korea. *International Journal of Food Microbiology*, vol. 134, pp. 196-200, 2009.

[19] R. A. Pollack, L. Findlay, W. Mondschein and R. R. Modesto. *Laboratory Exercises in Microbiology*. John Wiley and Sons, Hoboken, 2018.

[20] M. P. MacWilliams. *Indole Test Protocol*. American Society for Microbiology, Washington, D.C, 2012.

[21] S. McDevitt. *Methyl Red and Voges-Proskauer Test Protocols*. vol. 8, American Society for Microbiology, Washington, D.C, 2009.

[22] K. Reiner. *Catalse Test Protocol*. American Society For Microbiology, Washington, D.C, 2014.

[23] M. Ligozzi, C. Bernini, M. G. Bonora, M. De Fatima, J. Zuliani and R. Fontana. Evaluation of the VITEK 2 system for identification and antimicrobial susceptibility testing of medically relevant gram-positive cocci. *Journal of clinical microbiology*, vol. 40, pp. 1681-1686, 2002.

[24] I. Espinosa, M. Báez, M. I. Percedo and S. Martínez. Evaluation of simplified DNA extraction methods for *Streptococcus suis* typing. *Revista de Salud Animal*, vol. 35, pp. 59-63, 2013.

[25] L. Lin, B. D. Ling and X. Z. Li. Distribution of the multidrug efflux pump genes, adeABC, adeDE and adeIJK, and class 1 integron genes in multiple-antimicrobial-resistant clinical isolates of *Acinetobacter baumannii-Acinetobacter calcoaceticus* complex. *International Journal of Antimicrobial Agents*, vol. 33, pp. 27-32, 2009.

[26] M. Yavzori, N. Porath, O. Ochana, R. Dagan, R. Orni-Wasserlauf and D. Cohen. Detection of enterotoxigenic *Escherichia coli* in stool specimens by polymerase chain reaction. *Diagnostic Microbiology and Infectious Disease*, vol. 31, pp. 503-509, 1998.

[27] C. Rodas, V. Iniguez, F. Qadri, G. Wiklund, A. M. Svennerholm and A. Sjöling. Development of multiplex PCR assays for detection of enterotoxigenic *Escherichia coli* colonization factors and toxins. *Journal of Clinical Microbiology*, vol. 47, pp. 1218-1220, 2009.

[28] S. C. Pawar, A, T. A. P. Devi, C. Setti, R. Gajula, S. Srikanth, S. Kalyan. Molecular evolution of pathogenic bacteria based on rrsA gene. *Journal of Biotechnology and Biomaterials*, vol.2, pp.12-18, 2018.

[29] R. Mulchandani, F. Massebo, F. Bocho, C. L. Jeffries, T. Walker and L. A. Messenger. A community-level investigation following a yellow fever virus outbreak in South Omo Zone, South-West Ethiopia. *Peer Journal*, vol. 7, p. e6466, 2019.

[30] H. K. Hasan, N. A. Yassin and S. H. Eassa. Bacteriological and molecular characterization of diarrheagenic *Escherichia coli* pathotypes from children in Duhok City, Iraq. *Science Journal of University of Zakho*, vol. 8, pp. 52-57, 2020.

[31] T. W. Shatub, N. A. H. Jafar and A. K. Krekor Melconian. Detection of diarrheagenic *E.coli* among children under 5's age in Tikrit city of Iraq by using single multiplex PCR technique. *Plant Archives*, vol. 21, pp.1230-1237.16-118, 2021.

[32] Z. K. Khalil. Isolation and identification of different diarrheagenic (DEC) *Escherichia coli* pathotypes from children under five years old in Baghdad. *Iraqi Journal of Medical Sciences*, vol. 28, pp. 126-132, 2015.

[33] Z. K. Abdul-hussein, R. H. Raheema and A. I. Inssaf. Microbiology. Molecular diagnosis of diarrheagenic *E. coli* infections among the pediatric patients in Wasit Province, Iraq. *Journal of Pure and Applied Microbiology*, vol. 12, pp. 2229-2241, 2018.

[34] E. Abbasi, M. Mondanizadeh, A. van Belkum and E. J. I. Ghaznavi-Rad. Multi-drug-resistant diarrheagenic *Escherichia coli* pathotypes in pediatric patients with gastroenteritis from central Iran. *Infection and Drug Resistance*, vol. 13, p. 1387, 2020.

[35] A. Emami, N. Pirbonyeh, F. Javanmardi, A. Bazargani, A.Moattari, A. Keshavarzi, *et al*. Molecular diversity survey on diarrheagenic *Escherichia coli* isolates among children with gastroenteritis in Fars, Iran. *Future Microbiology*, vol. 16, pp. 1309-1318, 2021.

[36] H. K. Saka, N. T. Dabo, B. Muhammad, S. García-Soto, M. Ugarte-Ruiz and J. Alvarez. Diarrheagenic *Escherichia coli* pathotypes from children younger than 5 years in Kano State, Nigeria. *Frontiers in Public Health*, vol. 7, p. 348, 2019.

[37] G. Shahbazi, M. A. Rezaee, F. Nikkhahi, S. Ebrahimzadeh and F. Hemmati. Characteristics of diarrheagenic *Escherichia coli* pathotypes among children under the age of 10 years with acute diarrhea. Asian J Med Sci, vol. 25, p. 101318, 2021.

[38] G. Ochien and L. Atieno. Prevalence of Enterotoxigenic *Escherichia coli* Among Children Under Five Years in Siaya County, Western Kenya. Maseno University, Kenya, 2021.

[39] O. Lukjancenko, T. M. Wassenaar and D. W. Ussery. Comparison of 61 sequenced *Escherichia coli* Genomes. *Microbial Ecology*, vol. 60, pp. 708-720, 2010.

[40] S. K. Arif and L. I. F. Salih. Identification of different categories of diarrheagenic *Escherichia coli* in stool samples by using multiplex PCR technique. Asian J Med Sci, vol. 2, pp. 237-243, 2010.

[41] C. I. C. Ifeanyi, N. F. Ikeneche, B. E. Bassey, N. Al-Gallas, A. A. Casmir and I. R. Nnennaya. Characterization of toxins and colonization factors of enterotoxigenic *Escherichia coli* isolates from children with acute Diarrhea in Abuja, Nigeria. *Jundishapur Journal Of Microbiology*, vol. 11, p. e64269, 2018.

[42] S. Eybpoosh, S. Mostaan, M. M. Gouya, H. Masoumi-Asl, P. Owlia, B. Eshrati, *et al*. Frequency of five *Escherichia coli* pathotypes in Iranian adults and children with acute diarrhea. *PLoS One*, vol. 16, p. e0245470, 2021.

[43] S. Zheng, F. Yu, X. Chen, D. Cui, Y. Cheng, G. Xie, *et al*. Enteropathogens in children less than 5 years of age with acute diarrhea: A 5-year surveillance study in the Southeast Coast of China. *BMC infectious Diseases*, vol. 16, pp. 434-434, 2016.

[44] E. Kipkirui, M. Koech, A. Ombogo, R. Kirera, J. Ndonye, N. Kipkemoi, *et al*. Molecular characterization of enterotoxigenic *Escherichia coli* toxins and colonization factors in children under five years with acute diarrhea attending Kisii Teaching and Referral Hospital, Kenya. *Tropical Diseases Travel Medicine and Vaccines*, vol. 7, pp. 1-7, 2021.

[45] A. M. Svennerholm. From cholera to enterotoxigenic *Escherichia coli* (ETEC) vaccine development. *Indian Journal of Medical Research*, vol. 133, p. 188, 2011.

[46] S. Nazarian, S. L. M. Gargari, I. Rasooli, M. Alerasol, S. Bagheri and S. D. Alipoor. Prevalent phenotypic and genotypic profile of enterotoxigenic *Escherichia coli* among Iranian children. *Japanese Journal of Infectious Diseases*, vol. 67, pp. 78-85, 2014.

[47] H. Zeighami, F. Haghi, F. Hajiahmadi, M. Kashefiyeh and M. Memariani. Multi-drug-resistant enterotoxigenic and enterohemorrhagic *Escherichia coli* isolated from children with diarrhea. *Journal of Chemotherapy*, vol. 27, pp. 152-155, 2015.

[48] S. Gupta, J. Keck, P. K. Ram, J. A. Crump, M. A. Miller and E. D. Mintz. Part III. Analysis of data gaps pertaining to enterotoxigenic *Escherichia coli* infections in low and medium human development index countries, 1984-2005. *Epidemiology and Infection*, vol. 136, pp. 721-738, 2008.

[49] H. Alizade, R. Ghanbarpour and M. R. Aflatoonian. Molecular study on diarrheagenic *Escherichia coli* pathotypes isolated from under 5 years old children in southeast of Iran. *Asian Pacific Journal of Tropical Disease*, vol. 4, pp. S813-S817, 2014.

[50] E. Joffré, A. von Mentzer, A. M. Svennerholm and Å. Sjöling. Identification of new heat-stable (STa) enterotoxin allele variants produced by human enterotoxigenic *Escherichia coli* (ETEC). *International Journal of Medical Microbiology*, vol. 306, pp. 586-594, 2016.

[51] J. M. Fleckenstein, P. R. Hardwidge, G. P. Munson, D. A. Rasko, H. Sommerfelt and H. Steinsland. Molecular mechanisms of enterotoxigenic *Escherichia coli* infection. *Microbes and Infection*, vol. 12, pp. 89-98, 2010.

[52] O. Torres, W. González, O. Lemus, R. A. Pratdesaba, J. A. Matute, G. Wiklund, *et al*. Toxins and virulence factors of enterotoxigenic *Escherichia coli* associated with strains isolated from indigenous children and international visitors to a rural community in Guatemala. *Epidemiology and Infection*, vol. 143, pp. 1662-1671, 2015.

[53] S. X. Zhou, L. P. Wang, M. Y. Liu, H. Y. Zhang, Q. B. Lu, L. S. Shi, *et al*. Characteristics of diarrheagenic *Escherichia coli* among patients with acute diarrhea in China, 2009-2018. *Journal of Infection*, vol. 83, pp. 424-432, 2021.

[54] M. Paredes-Paredes, P. C. Okhuysen, J. Flores, J. A Mohamed, R. S Padda, A. Gonzalez-Estrada, *et al*. Seasonality of diarrheagenic *Escherichia coli* pathotypes in the US students acquiring diarrhea in Mexico. *Journal of Travel Medicine*, vol. 18, pp. 121-125, 2011.

[55] L. F. Peruski Jr., B. A. Kay, R. A. El-Yazeed, S. H. El-Etr, A. Cravioto, T. F. Wierzba, *et al*. Phenotypic diversity of enterotoxigenic *Escherichia coli* strains from a community-based study of pediatric diarrhea in periurban Egypt. *Journal of Clinical Microbiology*, vol. 37, pp. 2974-2978, 1999.

[56] H. I. Shaheen, S. B. Khalil, M. R. Rao, R. A. Elyazeed, T. F. Wierzba, L. F. Peruski Jr, *et al*. Phenotypic profiles of enterotoxigenic *Escherichia coli* associated with early childhood diarrhea in rural Egypt. *Journal of Clinical Microbiology*, vol. 42, pp. 5588-5595, 2004.

[57] D. G. Evans, D. J. Evans Jr and W. J. I. Tjoa. Hemagglutination of human group A erythrocytes by enterotoxigenic *Escherichia coli* isolated from adults with diarrhea: Correlation with colonization factor. *Infection and Immunity*, vol. 18, pp. 330-337, 1977.

[58] M. G. Jobling and R. K. Holmes. Type II heat-labile enterotoxins from 50 diverse *Escherichia coli* isolates belong almost exclusively to the LT-IIc family and may be prophage encode. *PLoS One*. vol. 7, p. e29898, 2012.

[59] F. Qadri, S. K. Das, A. S. Faruque, G. J. Fuchs, M. J. Albert, R. B. Sack, *et al*. Prevalence of toxin types and colonization factors in enterotoxigenic *Escherichia coli* isolated during a 2-year period from diarrheal patients in Banglades. *Journal of Clinical Microbiology*, vol. 38, pp. 27-31, 2000.

[60] B. Juma, P. Waiyaki, W. Bulimo, E. Wurapa, M. Mutugi and S. Kariuki. Molecular detection of Enterotoxigenic *Escherichia coli* surface antigens from patients in Machakos District Hospital, Kenya. *ECAMJ*, vol. 1, p. 62-68, 2014.

[61] M. R. C. Nunes, F. Penna, R. Franco, E. Mendes and P. Magalhaes. Enterotoxigenic *Escherichia coli* in children with acute diarrhoea and controls in Teresina/PI, Brazil: distribution of enterotoxin and colonization factor genes. *Journal of Applied Micorbiology*, vol. 111, pp. 224-232, 2011.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Photosynthetic Pigments and Stomata Characteristics of Cowpea (*Vigna sinensis* savi) under the Effect of X-Ray Radiation

Ikbal Muhammed Albarzinji[1], Arol Muhsen Anwar[1], Hawbash Hamadamin Karim[2], Mohammed Othman Ahmed[3]

[1]Department of Biology, Faculty of Science and Health, Koya University, Koya KOY45, Kurdistan Region - F.R. Iraq,
[2]Department of Physics, Faculty of Science and Health, Koya University, Koya KOY45, Kurdistan Region - F.R. Iraq,
[3]Department of Horticulture, College of Agricultural Engineering Sciences, University of Raparin, Kurdistan Region- F.R. Iraq

ABSTRACT

This study was conducted in the field and laboratories of the faculty of science and health-Koya university by exposing the seeds of cowpea plant (*Vigna sinensis* Savi) var. California black-eye to X-ray radiation in two different locations (In target or 30 cm out of target) inside the radiation chamber, for four different exposure times (0, 5, 10, or 20 min), to study the effect on some characteristics of seedling components. Results show that the exposure location to X-ray had non-significant effects on cowpea leaves content of photosynthetic pigments, whereas each of time of exposure with interaction between location and time of exposure had significant effects on chlorophyll a, total chlorophylls, and total carotenoids pigments. Regarding the X-ray effects on stomata characteristics, the results detect that there were non-significant differences between the location of exposure on stomata number on abaxial leaves surfaces and stomata length on adaxial leaves surfaces, whereas a significant effects on number of stomata on the adaxial leaves surfaces, abaxial stomata length, abaxial, and adaxial stomata width were detect. Exposing cowpea seeds to X-ray radiation in the target of the radiation source increased significantly stem and leave dry matter percent compared with the one out of the target location, whereas increasing the time of exposure decreased the percent of dry matter of stem and leaves. It is concluded that exposing cowpea seeds to X-ray leads to changes in photosynthetic pigments, stomata characteristics, and plant dry matter content.

Index Terms: *Vigna sinensis* savi, X-ray Radiation, Pigments, Stomata Traits

1. INTRODUCTION

In the field of agriculture, many practices particularly the using of chemicals are applied for improving crops quality and quantity, however, although their positive effects, these applications are not empty of undesirable effects on environment, public health, and plant growth. Using modern biotechnological approaches, including, electricity current, laser, magnetic field, high voltage, ultraviolet and radiation with gamma or X-ray on different plants material are gaining interest to develop plants growth and yield, and characterized by cheapness and safety on health and environment, therefore the scientists try to make this century a biophysical century, where most of the physical factors depend on increasing energy balance and increase material transport through membranes for improving the growth and the development of crops [1]-[3].

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp58-64

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Ikbal Muhammed Albarzinji, Department of Biology, Faculty of Science and Health, Koya University, Koya KOY45, Kurdistan Region - F.R. Iraq. E-mail: ikbal.tahir@koyauniversity.org

Received: 16-07-2022 Accepted: 07-09-2022 Published: 24-09-2022

Ionizing radiations are those have wavelengths <100 nm [4]. These radiations are charged high-energy particles (high-energy photons and electrons). Two types of ionizing radiations there are: Gamma radiations and X-rays, the first is emitted from inside the nucleus, whereas X-ray is radiated from outside the nucleus [5]. There are many applications of X-ray radiation in different fields of plant studies, for example Panchal *et al.* [6] used X-ray for imaging of inner features of a seed sample to identify unseen defects or contaminants. Other studies were conducted to investigate the effects of X irradiation on physiological characteristics of different plants, such; Rezk *et al.* [7] found that low dose of X-ray 5 Gray (Gy) caused increasing in all morphological criteria, total photosynthesis pigments, enzymatic and non-enzymatic antioxidants significantly in two genotypes of okra plants as compared with control treatments, while the doses (higher than 5 Gy) caused a considerable decreased in the studied parameters. Similarly, Singh [8] study shows promoting in chlorophyll development for 60 s X-ray pretreated as it compared to 90 and 120 s pre-treatment for seeds of *Cicer arietinum*, *Vigna radiata*, *Vigna mungo* and *Vicia faba* plants. Dhamgaye *et al.* [9] irradiated seeds of *Phaseolus vulgaris* cv. Rajmah using Synchroton X-ray Beam at 0.5–10 Gy, the overall growth of 10 days old seedlings raised from irradiated seeds was substantially reduced at irradiation doses of 2 and 5 Gy. Same authors Dhamgaye *et al.* [10] irradiated seeds of *P. vulgaris* cv. Rajmah using Synchrotron X Ray at doses of 1, 10, and 20 Gray where, the percent of relative water and protein content was significantly decreased at 10 and 20 Gy dose in 4–8 days old seedling, and a decrease in photosynthesis pigments chlorophyll and carotenoids content is observed in shoot tissue when 1 and 10 Gy where used. Mortazavi *et al.* [11] accelerated the growth of newly grown plants of *P. vulgaris* (Pinto) by irradiated them with X-rays for 6 days. Arena *et al.* [12] found that exposure of dwarf bean (*P. vulgaris* L.) plants to different doses of X-rays (0.3, 10, 50, and 100 Gy) showed that young leaves exhibited a reduction of area and an increase in specific mass and dry matter content. At higher doses of X-rays (50 and 100 Gy) total chlorophyll (a+b) and carotenoid (xanthophylls + carotenoids) content were significantly lower ($P < 0.01$) compared to lower doses and in control leaves. Significant reduction in transpiration was detected in *V. faba* irradiated by X-ray, this reduction was associated with inhibition of stomatal opening from the 9th to 16th day after irradiation. The osmotic pressure of epidermal cells in irradiated plants appeared to be slightly higher than that of epidermal cells of non-irradiated plants. However, the slight osmotic pressure changes of epidermal cells in irradiated plants did not appear to be a major factor contributing to inhibition of stomatal opening in irradiated plants under the growth conditions of the experiments [13].

The aim of this work was to investigate the effects of seed exposure to X-rays on some of the physiological properties of emerged cowpea plants, because these changes has subsequent effects on the photosynthetic activity and cause a direct effect on the agronomic features of the plant.

2. MATERIALS AND METHODS

2.1. Plant Materials and Studied Characteristics

This work was conducted in the Department of Biology/Koya University, Erbil-Iraq. The seeds of cowpea plant (*Vigna sinensis Savi*) var. California black-eye were exposed to a single dose of X-ray radiation by the XRD tube (from the Company of PANalytical B.V. Lelyweg1, the Netherlands) where the highest radiation level was less than 1 Sieverts/h measured at the tube surface. 20 seeds for each experimental unit were putted in the device source to exposed to X-ray at the advanced physics laboratory in physic department at same faculty. The experiment was conducted in complete randomize design (CRD) where the location of exposure considers as the first factor by exposure the seeds to X-rays either in the target point of the device or 30 cm from the target point in the base of the device chamber, whereas the times of exposure 0, 5, 10, or 20 min were considered as the second factor, where the time zero is considered as the control treatment used for each location.

After seeds were exposed to X-ray they planted in 5 kg. soil pots, because of an initial increase in photosynthesis rate during leaf expansion and followed by a decrease on maturation [14], at the end of the vegetative growth stage, fourth leaf of five plants from each experimental unit were taken, and the photosynthetic pigments chlorophylls a, chlorophyll b, total chlorophylls and total carotenoids were estimated as it mentioned in Lichtenthaler and Wellburn [15] were leaf material was collected and mixture ratio was 50 ml 80% acetone: 1 g leaves sample. Samples were grinded by mortar and pestle and filtered by filter paper, then extracts were placed in a 25 ml dark glass vial to avoid evaporation and photo-oxidation of pigments, after that the absorbance of the extract were measured by spectrophotometer at wave lengths 663, 646, and 470 nm. Each of chlorophyll a, b, and total carotenoids were estimated as follows:

$$\text{Chlorophyll a} = (12.21*A663) - (2.81*A646)$$

$$\text{Chlorophyll b} = (20.13*A646) - (5.03*A663)$$

$$\text{Total carotenoids} = (1000*A470 - 3.27*\text{Chl a} - 104*\text{Chl b})/229$$

Where, A is Absorbance, chl. a = chlorophyll a (mg/L) and chl. b = chlorophyll b (mg/L).

For converting the concentration from mg/L to mg/g fresh weight, each value multiplied by (extraction volume/sample weight *1000), and total chlorophyll calculated from the summation of each chlorophyll a and chlorophyll b.

Total chlorophyll was determined by collecting each of chlorophyll a and chlorophyll b (10).

For stomata study, the lasting impressions method [16] was used. In this method, about one square centimeter of leaves surfaces was painted by a clear nail polish. After the nail polish was dried they were taped by a clear cellophane tape, and peels it out. The leaf impressions taped on slides and labeled as adaxial and abaxial surfaces then examined under ×40 by light microscope (DM 300, Leica Microsystems, China). Numbers of appeared stomata on lens field were counted for all adaxial and abaxial leaves surfaces. Stomata guard cells length and width of adaxial and abaxial leaves surfaces were calculated in micrometer (μm) with scaled ocular lens.

Because of the important of the percent of dry matter content as a result of the photosynthetic activity it determined for each of stem and leaves by dividing the stem or leaves dry weight by the stem or leaves fresh weight multiplying by 100 as it reported by Al-Sahaf [17].

2.2. The Statistical Analysis

The statistical analysis of the study conducted as a factorial experiment performed as CRD in three replications, analysis of variance was used for calculating the differences among each factor treatments and their interactions by using the SAS software. The test of Duncan's multiple comparison was used to estimate the main effects of treatments which were differ when the F-value was significant at $P \leq 0.05$ [18].

3. RESULTS AND DISCUSSION

From Table 1 results it is shown that location of exposure to X-ray had non-significant ($P > 0.05$) effects on leaves content of photosynthesis pigments of cowpea plant, whereas the time of exposure led to a significant ($P \leq 0.05$) effects on chlorophyll a and total chlorophylls, were the seeds that exposed to X-ray for 10 min increased chlorophyll a and total chlorophylls significantly ($P \leq 0.05$) to 2.64 and 5.42 mg/g fresh weight comparing to other exposure times. Results of interactions between locations and time of exposure revealed that exposure for 10 min out of target increased the content of chlorophyll a significantly to 3.01 mg/g fresh weight compared to other interaction treatments, and to 3.14 and 6.15 mg/g fresh weight for each of chlorophyll b and total chlorophylls compared with 5 min exposure out of target only, whereas same interaction increased total carotenoids content significantly to 1.15 mg/g fresh weight compared to 0.94 mg/g fresh weight for 20 min exposure out the target interaction treatment only. In general, ionizing radiation may have different effects on plant metabolism, growth and

TABLE 1: Effects of location and time of exposure cowpea seed to X-radiation, and their interactions on chlorophyll a, b, total chlorophylls and total carotenoids

Treatments	Chlorophyll (mg/g fresh weight) a	Chlorophyll (mg/g fresh weight) b	Total chlorophylls (mg/g fresh weight)	Total carotenoids (mg/g fresh weight)
Location of exposure				
In target (L1)	2.23^{a}	2.33^{a}	4.56^{a}	1.05^{a}
Out of target (L2)	2.32^{a}	2.38^{a}	4.70^{a}	1.04^{a}
Time of exposure (min)				
T0	2.20^{b}	2.27^{a}	4.47^{ab}	1.07^{a}
T5	2.11^{b}	1.96^{a}	4.07^{b}	1.01^{a}
T10	2.64^{a}	2.78^{a}	5.42^{a}	1.09^{a}
T20	2.16^{b}	2.40^{a}	4.56^{ab}	0.99^{a}
Interactions between location and exposure time				
L1×T0	2.20^{b}	2.27^{ab}	4.47^{ab}	1.07^{ab}
L1×T5	2.27^{b}	2.27^{ab}	4.54^{ab}	1.02^{ab}
L1×T10	2.27^{b}	2.42^{ab}	4.70^{ab}	1.03^{ab}
L1×T20	2.18^{b}	2.34^{ab}	4.52^{ab}	1.05^{ab}
L2×T0	2.20^{b}	2.27^{ab}	4.47^{ab}	1.07^{ab}
L2×T5	1.96^{b}	1.65^{b}	3.60^{b}	0.99^{ab}
L2×T10	3.01^{a}	3.14^{a}	6.15^{a}	1.15^{a}
L2×T20	2.13^{b}	2.46^{ab}	4.59^{ab}	0.94^{b}

Means that followed by same letters within column are differ non-significantly at $P \leq 5\%$ according to the Duncan multiple range test

reproduction, depending on radiation dose, plant species, developmental stage, and physiological traits [12]. Our results disagree with the Al-Enezi and Al-Khayri [19] results that suggested that photosynthesis pigments chlorophyll a and carotenoids are more sensitive to X-ray than chlorophyll b, whereas we found that chlorophyll b and total carotenoids were less sensitive to X-irradiation compared to chlorophyll a and total chlorophylls. Changes in photosynthetic pigments were studied by Arena *et al.* [12] whom confirmed that the decrease in the levels of X-ray (0.3 Gy) caused an increase in photosynthetic pigments in bean plants, whereas the high levels (50 and 100 Gy) caused a decrease in these pigments, these findings also agree with that of Rezk *et al.* [7] which recorded in two okra genotypes leaves, where the content of photosynthetic pigment improved significantly with increasing the doses of X-ray to 5 Gy comparing with untreated plants, also more increase in the radiation doses, encourage the reduction in photosynthetic pigments compared to the control plants. Changes in chlorophyll content as a response to X-ray is either toward an increase or a decrease direction, the increase may due to the increase in chlorophyll biosynthesis and/or delaying its degradation [20], whereas the decrease may due to pigment breakdown due to increase of reactive oxygen species [21] and changes in the chloroplast such chloroplast swelling, thylakoid dilation, and breakdown of chloroplast outer membrane [22].

Regarding X-radiation effects on the stomata characteristics it was shown that there were non-significant ($P > 0.05$) differences between the location of exposure on number of stomata on abaxial leaves surfaces and stomata length on adaxial surfaces of leaves (Table 2 and Figs. 1 and 2). The seeds exposed directly to the source of X-ray (in target) decreased number of stomata on the adaxial leaves surfaces to 148.33 stomata/mm^2, whereas abaxial stomata length increased to 11.08 micrometer and abaxial with adaxial stomata width also increased significantly to 7.00 and 7.58 micrometer, respectively, compared to 180.00 stomata/mm^2, 9.92, 5.42, and 5.50 micrometer for plants out of target. 10 min of seed exposure to X-ray increased stomata number on both abaxial and adaxial leaves surfaces compared to other exposure times except the control treatment. Exposure time had non-significant ($P > 0.05$) effect on stomata length and width on abaxial leaves surfaces, whereas increasing time of exposure to 20 min increased the stomata length significantly ($P \leq 0.05$) compared to 5 and 10 min only, whereas it increased the stomata width significantly compared to all other treatments. From the results of interaction within location and time of exposure, it was clear from the results (Table 2), that treating seeds for 10 min in the X-ray target had the more significant effects for abaxial leaves surfaces in increasing stomata number to 540.00 stomata/mm^2, and the stomata length and width to 11.67 and 8.00

TABLE 2: Effects of seeds exposure to X-radiation on some characteristics of cowpea (*Vigna sinensis* Savi) plants stomata

Treatments	Stomata number/mm^2		Stomata length (micrometer)		Stomata width (micrometer)	
	Abaxial leaves surface	Adaxial leaves surface	Abaxial leaves surface	Adaxial leaves surface	Abaxial leaves surface	Adaxial leaves surface
Location of exposure						
In target (L1)	455.00^{a}	148.33^{b}	11.08^{a}	10.58^{a}	7.00^{a}	7.58^{a}
Out of target (L2)	455.83^{a}	180.00^{a}	9.92^{b}	11.67^{a}	5.42^{b}	5.50^{b}
Time of exposure (min)						
T0	526.67^{a}	176.67^{ab}	10.67^{a}	12.33^{a}	6.00^{a}	6.00^{bc}
T5	398.33^{b}	146.67^{bc}	10.33^{a}	9.67^{b}	6.17^{a}	5.17^{c}
T10	536.67^{a}	201.67^{a}	10.17^{a}	10.17^{b}	6.00^{a}	6.67^{b}
T20	360.00^{b}	131.67^{c}	10.83^{a}	12.33^{a}	6.67^{a}	8.33^{a}
Interactions between location and exposure time						
L1×T0	526.67^{ab}	176.67^{bc}	10.67^{ab}	12.33^{ab}	6.00^{abc}	6.00^{bc}
L1×T5	403.33^{bc}	143.33^{bc}	10.67^{ab}	10.33^{bcd}	7.33^{ab}	6.00^{bc}
L1×T10	540.00^{a}	156.67^{bc}	11.67^{a}	8.33d	8.00^{a}	8.67^{a}
L1×T20	350.00^{c}	116.67^{c}	11.33^{a}	11.33^{abc}	6.67^{ab}	9.67^{a}
L2×T0	526.67^{ab}	176.67^{bc}	10.67^{ab}	12.33^{ab}	6.00^{abc}	6.00^{bc}
L2×T5	393.33^{c}	150.00^{bc}	10.00^{ab}	9.00cd	5.00^{bc}	4.33^{c}
L2×T10	533.33^{a}	246.67^{a}	8.67b	12.00^{ab}	4.00^{c}	4.67^{cd}
L2×T20	370.00^{c}	146.67^{bc}	10.33^{ab}	13.33^{a}	6.67^{ab}	7.00^{b}

Means that followed by same letters within column are differ non-significantly at $P \leq 5\%$ according to the Duncan multiple range test

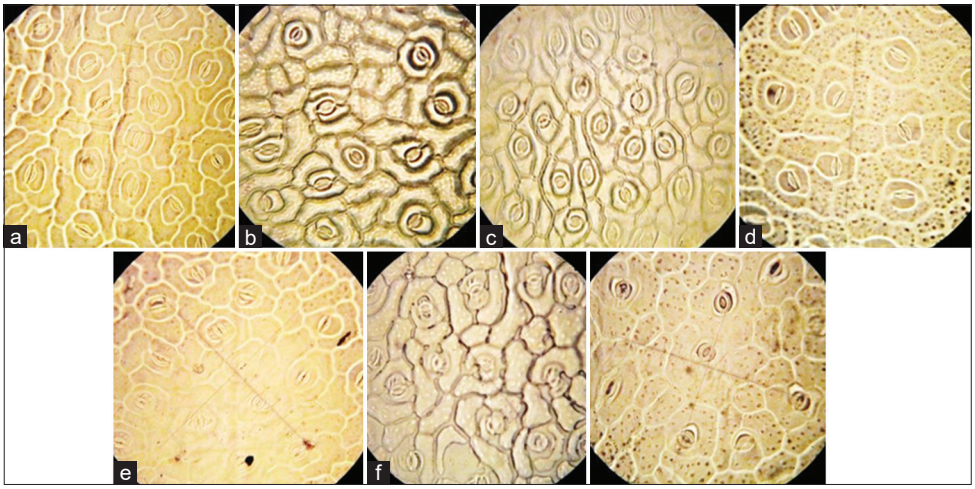


Fig. 1. Lower (abaxial) leaves surfaces of *Vigna sinensis* Savi showing stomata at ×400 for (a) the control, (b) in-target −5 min. (c) in-target −10 min. (d) in-target −20 min. (e) out of target −5 min. (f) out of target −10 min., and (g) out of target −20 min.

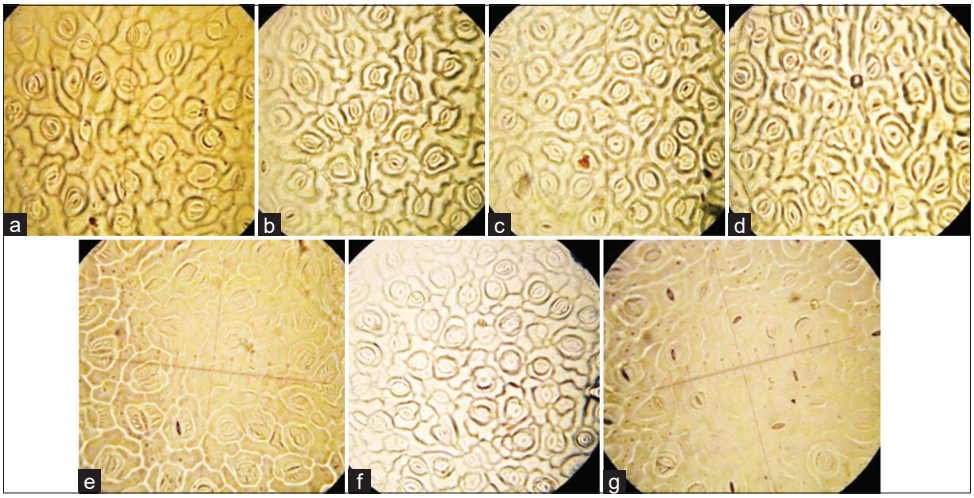


Fig. 2. Upper (adaxial) leaves surfaces of *Vigna sinensis* Savi showing stomata at ×400 for (a) the control, (b) in-target −5 min. (c) in-target −10 min. (d) in-target −20 min. (e) out of target −5 min. (f) out of target −10 min. and (g) out of target −20 min.

micrometer, respectively, in coincides with the treatment 20 min exposure time in the target of radiation source for adaxial leaves surfaces which increased stomata width to 8.67 and 9.67 micrometer, respectively. The present observations showed changes in stomata characteristics under X-ray radiation compared with that not treated. These changes in stomata dimensions under X-ray may due to change in osmotic pressure of epidermal cells which prevent the development of sufficient osmotic pressure in guard cells to open to the same extent as occurs in non-irradiated plants, so the average stomatal opening of X-ray irradiated plants was significantly less compared to non-irradiated plants [13]. Stomatal aperture depends on the genotype of plants and is regulated by many internal and external factors [23].

From Table 3 results, it is shown that exposure seeds to X-ray in target source increases significantly each of stem and leaves dry matter percent to 10.75 and 14.00% compared to that is out of target location (9.13 and 11.88%), respectively, which agrees with Al-Enezi and Al-Khayri [24] whom found an increase in fresh and dry weights of date palm (*Phoenix dactylifera* L.) leaf tissues with increasing the X irradiation dose from 0 to 1500 rad, it also agrees with the results of Arena *et al*. [12] whom found that the high dose of X-rays (50 Gy) increased significantly ($P < 0.001$) leaf dry matter content in faba been young leaves compared to the control leaves. Regarding the time of exposure 5 min exposure to X-ray increased the percent of stem and leaves dry matter content significantly ($P \leq 0.05$) to 13.75 and

TABLE 3: Effects of location and time of exposure cowpea seed to X-radiation, and their interactions in stem and leaf dry matter

Treatments	Stem dry matter (%)	Leaves dry matter (%)
Location of exposure		
In target (L1)	10.75[a]	14.00[a]
Out of target (L2)	9.13[b]	11.88[b]
Time of exposure (min)		
T0	8.50[b]	12.00[bc]
T5	13.75[a]	15.75[a]
T10	9.50[b]	13.50[b]
T20	8.00[b]	10.50[c]
Interactions between location and exposure time		
L1×T0	8.50[cd]	12.00[c]
L1×T5	15.00[a]	17.00[a]
L1×T10	10.00[c]	14.00[bc]
L1×T20	9.50[c]	13.00[bc]
L2×T0	8.50[cd]	12.00[c]
L2×T5	12.50[b]	14.50[b]
L2×T10	9.00[c]	13.00[bc]
L2×T20	6.50[d]	8.00[d]

Means that followed by same letters within column are differ non-significantly at $P \leq 5\%$ according to the Duncan multiple range test

15.75% compared to other treatments, whereas increasing the time of exposure to 20 min decreased the percent of stem and leaves dry matter significantly ($P \leq 0.05$) to 8.00% and 10.5%, respectively. Regarding the interactions between the location and time of exposure, the percent of stem and leaves content of dry matter increased significantly to 15.00% and 17.0% for plants emerged from seeds exposed to X-ray for 5 min on the source target, whereas the lowest values were recorded for seeds exposed to 20 min X-ray out of target. For the X-ray effects, it was seen that the shortest time records led to the highest significant increase, which can be concluded that it likes the effects of low doses of X-radiation which encourage cellular activities and growth whereas higher doses may cause chromosomal abnormalities [25]. Hence, higher X-ray radiation exposure time effect on the growth of plants which reflects on stem and leaves percent of dry matter.

4. CONCLUSIONS

We can conclude that exposing cowpea seeds to X-ray radiation had stimulation effects regarding photosynthesis pigments and stomata characteristics either as increase or decrease responses according to the treatment. It was concluded that the location of the exposure had non-significant effects on photosynthetic pigments, whereas, it effects on stomata characteristics and dry matter content. Best exposure time differ according to the studied characteristics. More studies are recommended about effects of X-ray on wet seeds and seedling by different doses of radiation.

REFERENCES

1. G. Vasilevski. "Perspectives of the application of physiological methods in sustainable agriculture". *Bulgarian Journal for Plant Physiol, Special Issue*, vol. 3-4, pp. 179-186, 2003.
2. M. K. Al-Jebori and I. M. Al-Barzinji. "Exposing potato seed tuber to high voltage field I. Effects on growth and yield". *Journal of Iraqi Agricultural Sciences*, vol. 39, no. 2, pp. 1-11, 2008.
3. I. M. Al-Barzinji and M. K. Al-Jubouri. "Effects of exposing potato tuber seeds to UV radiation on growth, yield and yield quality". *Research and Reviews: Journal of Botany*, vol. 5, no. 2, pp. 19-26, 2016.
4. K. H. Ng. "Non-Ionizing Radiations-Sources, Biological Effects, Emissions and Exposures". *Proceedings of the International Conference on Non-Ionizing Radiation at UNITEN (ICNIR2003). Electromagnetic Fields and Our Health*, 2003.
5. Environmental Protection Agency. "Radiation: Facts, Risks and Realities". Environmental Protection Agency, Washington, D.C, United States, 2012. Available from: https://www.epa.gov/sites/default/files/2015-05/documents/402-k-10-008.pdf [Last accessed on 2022 Sep 23].
6. K. P. Panchal, N. R. Pandya, S. Albert and D. J. Gandhi. "A x-ray image analysis for assessment of forage seed quality". *International Journal of Plant, Animal and Environmental Sciences*, vol. 4, no. 4, pp. 103-109, 2014.
7. A. A. Rezk, J. M. Al-Khayri, A. M. Al-Bahrany, H. S. El-Beltagi and H. I. Mohamed. "X-ray irradiation changes germination and biochemical analysis of two genotypes of okra (*Hibiscus esculentus* L.)". *Journal of Radiation Research and Applied Sciences*, vol. 12, no. 1, pp. 393-402, 2019.
8. J. Singh. "Studies in bio-physics: Effect of electromagnetic field and x-rays on certain road side legume plants at Saharanpur". *International Journal of Scientific and Research Publications*, vol. 3, no. 12, pp. 1-9, 2013.
9. S. Dhamgaye, V. Dhamgaye and R. Gadre. "Growth retardation at different stages of bean seedlings developed from seeds exposed to synchrotron x-ray beam". *Advances in Biological Chemistry*, vol. 8, no. 2, pp. 29-35, 2018.
10. S. Dhamgaye, N. Gupta, A. Shrotriya, V. Dhamgaye and R. Gadre. "Biological effects of seed irradiation by synchrotron x-ray beam in young bean seedlings". *Advances in Biological Chemistry*, vol. 9, no. 2, pp. 88-97, 2019.
11. S. M. Mortazavi, L. A. Mehdi-Pour, S. Tanavardi, S. Mohammadi, S. Kazempour, S. Fatehi, B. Behnejad and H. Mozdarani. "The biopositive effects of diagnostic doses of X-rays on growth of *Phaseolus vulgaris* plant: A possibility of new physical fertilizers". *Asian Journal of Expermental Sciences*, vol. 20, no. 1, pp. 27-33, 2006.
12. C. Arena, V. De Micco and A. De Maio. "Growth alteration and leaf biochemical responses in *Phaseolus vulgaris* exposed to different doses of ionising radiation". *Plant Biology*, vol. 16, no. Suppl 1, pp. 194-202, 2014.
13. R. M. Roy. "Transpiration and stomatal opening of x-irradiated broad bean seedlings". *Radiation Botany*, vol. 14, no. 3, pp. 179-184, 1974.
14. C. Z. Jiang, S. R. Rodermel and R. M. Shibles. "Photosynthesis, rubisco activity and amount, and their regulation by transcription

in senescing soybeen leaves". *Plant Physiology*, vol. 101, no. 1, pp. 105-112, 1993.

15. K. Lichtenthaler and A. R. Wellburn. "Determination of total carotenoids and chlorophylls a and b of leaf extracts in different solvents". *Biochemical Society Transactions*, vol. 11, no. 5, pp. 591-592, 1983.
16. R. Priyanka and R. M. Mishra. "Effect of urban air pollution on epidermal traits of road side tree species, *Pongamia pinnata* (L.) Merr". *ISRO Journal of Environmental Science, Toxicology and Food Technology*, vol. 2, no. 6, pp. 2319-2402, 2013.
17. F. H. Al-Sahaf. "Applied Plant Nutrition. University of Baghdad. Ministry of Higher Education and Scientific Research". Dar al-Hikma Press, Iraq, p. 260, 1989.
18. A. H. Reza. "Design of Experiments for Agriculture and the Natural Sciences". 2nd ed. Chapman and Hall/CRC, New York, pp. 452, 2006.
19. N. A. Al-Enezi, and J.M. Al-Khayri. "Alterations of DNA, ions and photosynthetic pigments content in date palm seedlings induced by X-irradiation". *International Journal of Agricultural and Biology*, vol. 14, no. 3, pp. 329-336, 2012a.
20. A. A. Aly, R. W. Maraei, and S. Ayadi. "Some biochemical changes in two Egyptian bread wheat cultivars in response to gamma irradiation and salt stress". *Bulgarian Journal of Agricultural Science*, vol. 24, no. 1, pp. 50-59, 2018.
21. L. R. Dartnell, M. C. Storrie-Lombardi, C. W. Mullineaux, A. V. Ruban, G. Wright, A. D. Griffiths, J. P. Muller and J. M. Ward. "Degradation of cyanobacterial biosignatures by ionizing radiation". *Astrobiology*, vol. 11, no. 10, pp. 997-1016, 2011.
22. H. H. Latif and H. I. Mohamed. "Exogenous applications of moringa leaf extract effect on retrotransposon, ultrastructural and biochemical contents of common bean plants under environmental stresses". *South African Journal of Botany*, vol. 106, pp. 221-231, 2016.
23. L. Taiz and E. Zeiger. "Plant Physiology". 3rd ed. Sinauer Associates Publications, Sunderland, Massachusetts, p. 690, 2002.
24. N. A. Al-Enezi and J. M. Al-Khayri. "Effect of X-irradiation on proline accumulation, growth and water content of date palm (*Phoenix dactylifera* L.) seedlings". *Journal of Biological Sciences*, vol. 12, no. 3, pp. 146-153, 2012b.
25. D. O. Kehinde, K. O. Ogunwenmo, B. Ajeniya, A. A. Ogunowo and A. O. Onigbinde. "Effects of X-ray irradiation on growth physiology of *Arachis hypogaea* (Var. Kampala)". *Chemistry International*, vol. 3, no. 3, pp. 296-300, 2017.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Performance Assessment of Teaching through Students Evaluations: A Case Study Applied at University of Anbar

Muzhir Shaban Al-Ani*

Department of Information Technology, University of Human Development, College of Science and Technology, Sulaymaniyah, Kurdistan Regional Government, Iraq

ABSTRACT

The methods of the assessment of faculty member in universities varied. Some of them depend on the assessment of the faculty member on the teaching and research burden. Others assessments focus on how much teaching is provided during the calendar year. There are some ways in which the faculty member's assessment depends on the qualitative aspect provided by the faculty member. In addition, some of them combine both quantitative and qualitative aspects of the necessities to reach an appropriate evaluation process. This research aimed to consider the student as an essential part in the educational science and it is clear that the role taken by the student is no less important than the other role. This research started, where a questionnaire was prepared which included several paragraphs to measure the extent to which the student can evaluate the professor. The fourth stage students were chosen as the study sample because the students at last stage in the university has a good level of thinking and preparation. That means these students have experience in understanding the level of teaching and the knowledge of teachers. The proposed approach tries to analyze the educational model for evaluation developed in the Computer Science Department of the University of Anbar for learning and teaching of computer science and related scientific disciplines. In general, it is clear that more than 50% of the tested sample are satisfied with the teaching process which indicated good results have been achieved.

Index Terms: Teaching Assessment, Teaching Evaluation, Evaluations via Students, Teaching Case Study

1. INTRODUCTION

It is important to introduce a brief overview of Al-Anbar University and the faculty of Computer Science and Information Technology, in which the sample has been implemented. Al-Anbar University was founded in 1987 and is located in the city of Ramadi and it is a public university. In addition to being the only university in the province, there is another private college established at the same time. At its inception, the university consisted of four simple faculties. The university then grew to include 20 colleges and six campuses. Before 2003, the students came to the university from all governorates of Iraq, then after 2003, due to the abnormal conditions in Iraq, most university students are residents of the same governorate. Most university students are from middle- and low-income classes in society and this applies to university members.

The Faculty of Computers and Information Technology was established in 1998 and has two departments: Computer Science and Information Systems. The focus of these two departments is to provide the state and private sector with cadres with experience and knowledge in the fields of

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp65-76 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Muzhir Shaban Al-Ani, Department of Information Technology, University of Human Development, College of Science and Technology, Sulaymaniyah, Kurdistan Regional Government, Iraq. E-mail: muzhir.al-ani@uhd.edu.iq

Received: 29-07-2022 Accepted: 06-09-2022 Published: 01-10-2022

computer science and information technology [1]. This faculty has been grown and expanded and has been able to open programs for the study of Master's, in addition to it has attracted professors with competence and experience in the field of specialization [2]. Students are admitted to this faculty according to the central admission system prepared by the Ministry of Higher Education and Scientific Research, noting that admission rates depend on that year, which is generally within the average of 80% [3]. The number of students admitted to the faculty annually up to 150 students and the total number of students in the faculty is about 500 students, in addition there are 40 faculty members (M.Sc. and Ph.D.) [4].

The progress of societies depends primarily on the progress of education, especially university education, which is the vital entity that feeds society with the human resources and expertise that society needs in the development and advancement [5]. This is primarily based on the efficiency and potential of the faculty members [6].

2. TEACHING AND EDUCATION

Study in most Arab countries (including Iraq) is divided into three main phases; primary study, undergraduate study and post graduate study and these studies complement each other [7]. The primary study is an important study that feeding the students with fundamental subjects and knowledge, this study including custody of children (age of 4 year), preschool (age of 5 year), primary school (age of 6–11 years), intermediate school (age of 12–14 years), and secondary school (age of 15–18 years) [8]. Bachelor's study focusing on teaching the students with specific knowledge according to their specialty [9]. This study usually takes 4 years, except for some disciplines that need 5 or 6 years [10]. Postgraduate study including Master of Science (2–3 years) and philosophy of doctor (3–4 years) [11].

This paper will be concentrated on the undergraduate study (university study). The university study is one of the most important stages of study because it provides students with skills and knowledge [12]. The number of universities in Iraq has steadily increased to be 30, in addition many of these universities complain of a lack of trained and experienced staff [13]. The reason is the migration of many experienced academic staff abroad due to the difficult circumstances in Iraq [14].

The educational process focuses on three basic tasks namely teaching, scientific research and community service [15], [16]. In the Arab world, the educational process focuses on the task of teaching only and neglecting the other tasks as most of the universities have forgotten the task of community service, in addition in the task of scientific research there is no mentioned support [17], [18]. Hence, it is clear that the universities in the Arab world have lost a lot of importance and have become a mere view to get the certificate only [19].

The number of universities in Iraq in the seventies was only six universities and was characterized by scientific compass and was accredited to all universities in the world [20], [21]. Where it exited thousands of scientists and geniuses whom had a distinctive role all over the world [22]. After this important role of the universities and after 50 years of that period of time, the development of education globally entered the world of modern technologies and global universities have a constructive role in the management of scientific research, but in Arab universities, including Iraq, still behind the development [23], [24].

Most officials of Iraqi universities talk about the quality of education, but the majority do not understand the meaning of quality [25]-[27]. Everyone understand the quality as the quantity of papers, data and forms that are mobilized to appear well before the highest official in the ministry [28]-[30]. To understand that quality, it is planning and vision of the future represented by acts and continuous work [31], [32]. This process begins with the highest official in the institution or the university and finish with a small employee [33], [34]. These activities must be in the service of the students, community, and the educational process [35], [36].

3. RELATED WORK

Many works are published related to teaching process assessment as mentioned below, and it is summarized in Table 1.

Kiersma *et al.* (2016) proposed and approach to identify and evaluate the evidence, processes and criteria used to select the recipients of the teaching awards. In addition, informed best practices to select the recipients of the teaching awards. A specific sample of AACP members and pharmacy students was invited to an online survey on the process for nominating and selecting winners of educational awards, as well as perceptions of best practices [37].

Barana and Marchisio (2016) analyzed the educational model for automated formative evaluation developed in

TABLE 1: Summarizing table of the related work

References #	year	Author	Sample	Method
[37]	2016	Kiersma *et al*.	On line survey AACP members and pharmacy students	process for nominating and selecting winners of educational awards
[38]	2016	Barana, *et al*.	Department of Mathematics of University of Turin	analyze an educational model for automated formative assessment
[39]	2017	Artés *et al*.	Spanish university system	relationship between research performance and teaching quality
[40]	2017	Alhija	Israeli students' through internet survey designed to measure students' conceptions	Examined the relationship between these conceptions and students' background characteristics.
[41]	2018	Wikander, *et al*.	medicine and nursing skills	student feedback and a cross-sectional, qualitative study exploring
[42]	2018	Nguyen, *et al*.	Teaching English to Speakers of Other Languages	semi-structured in-depth interviews, and classroom observations
[43]	2019	Cano-Moreno *et al*.	Project Management university subject, School of Engineering and Industrial Design of the Polytechnic University of Madrid	quantitative methodology for the assessment of a university subject
[44]	2020	Li *et al*.	Many nets are trained on the evaluation sets of students	Integrating way to synthesize the results of three sub-networks.
[45]	2021	Lohman	student evaluations of teaching (SET) from different sources.	full potential of human resources tools to support consistent evaluation of teaching remains unrealized
[46]	2022	Romero *et al*.	Higher education level, from different sources.	generate preventive actions to improve educational quality
	2022	My study	Computer science students at university of Anbar-Iraq	Face to face Teaching process evaluation via students feedback

the Mathematics Department of the University of Turin for the learning and teaching of mathematics and scientific disciplines. The model is provided by an automated scoring system that powered by the engine of an advanced computer environment, allows the creation of algorithmic variables and opens mathematical responses, recognized in all their equivalent forms. The results obtained are discussed by means of the application of the automated formative evaluation in several class experiments and the data on the satisfaction and criticisms issued also shown [38].

Artés *et al*. (2017) studied the relationship between the performance of research and the quality of teaching in the context of the Spanish university system. They examined whether there is a relationship between being an active researcher and the quality of teaching of university professors in Spain. They used a set of data from the University of Extremadura, which contains information on the evaluation of teaching and the conduct of research over a period of 10 years (2001-2002–2011-2012). The obtained results suggested that, on average, the teachers most involved in the research obtain better results in their educational evaluations [39].

Alhija (2017) explored student's conceptions of good teaching of and examined the relationship between these conceptions and the basic characteristics of the students. Data were collected through an online survey designed to measure students' conceptions of five dimensions of instruction related to achievement goals, long-term student development, teaching methods, student relationships, and evaluation. The results indicated that students believe that assessment is the most important of the five dimensions of instruction and that long-term student development is the least important. Only gender and field of study have made a significant difference in students' perceptions of good teaching. In addition, implications for the evaluation of teaching are discussed [40].

Wikander and Bouchoucha (2018) provided an overview of the process leading to the successful adaptation of structured objective clinical assessment to meet the requirements of a pre-taught nursing course through blended learning. This is important because many universities move their study program online or in combination, while little attention has been paid to the adaptation of the evaluation of simulated clinical skills. The objective is to identify the advantages and disadvantages of objective structured and peer-reviewed clinical evaluation and share recommendations for successful implementation [41].

Nguyen and Walkinshaw (2018) examined the extent of teaching English to speakers of other languages that training

TABLE 2: Questionnaire design

Basic paragraphs to measure the performance of the faculty member		1	2	3	4	5
1st	Possibilities of the faculty member in teaching, ranking and preparing the material					
1	The course study plan is distributed in the 1st week.					
2	The scientific course is presented in a clear, coherent, and systematic manner.					
3	Utilizes lecture time effectively.					
4	The extent to which the teacher is able to use and present the scientific material.					
5	The compatibility of the plan's vocabulary with what has already been taught.					
6	Commits to the dates of his lectures accurately.					
2nd	Contribution of scientific material in the educational achievement of students					
7	Students are encouraged to participate and express their views on the scientific subject.					
8	Shows interest in students' academic achievement in general.					
9	Deals with students with respect within the standards of the profession and ethics.					
10	Uses teaching methods that stimulate thinking and curiosity.					
11	In the presentation of the scientific material, illustrations and applied methods are used.					
12	Different methods of teaching used to suit the subject matter of science and the needs of students.					
13	Uses a clear and understandable language in the teaching of the scientific subject.					
3rd	Evaluation of the content of the scientific material					
14	The content of the examinations is consistent with the explanation of the vocabulary of the teaching plan of the course.					
15	Discusses with the students the correct answers to the questions included in the exam.					
16	Uses different methods for measuring student achievement and assessing their scores.					
17	Students are asked for more than one exam to determine the degrees and scores.					
18	Students are frequently assigned to tasks and assignments.					
19	Quizzes are used frequently for students.					
4th	Relationship between faculty member and students					
20	The teacher is committed to office hours and encourages students to use them.					
21	Responds to students in answering their questions.					
22	Accuracy and fairness in student assessment.					
23	Students are encouraged to use different references to scientific material.					
24	Encourage students to respectful attitudes, customs and ethics.					
25	Deals with students on the basis of the principle of equality.					

in English speaking countries of the inner circle has had an impact on the autonomy in the teaching of Vietnamese English teachers. Through an online survey, in-depth semi-structured interviews and observations in the classroom, the research explored the tensions felt by these teachers when they tried to exercise their autonomy after returning to their institution. This document has significant implications for a variety of stakeholders involved in the professional development of non-inner circle teachers trained in inner circle contexts [42].

Cano-Moreno *et al.* (2019) provided a quantitative method for evaluation the university teaching. Companies, students and professors are involved in this assessment. This method is realized via four matrices concatenation applied on project management course at school of engineering in Polytechnic University of Madrid. One of the big advantages of this study focusing on improving skills and knowledge of the selected subject [43].

Li *et al.* (2020) improved the quality of teaching process in academic institutions. This research introduced an efficient neural network approach in order to evaluate the quality of teaching process. Leaders, peers, and students are trained and evaluated leading to improve their performance. In addition, on line system approach was designed and implemented for teaching evaluation to provide suitable environment [44].

Lohman (2021) studied the teaching process evaluation through student's feedback. Evaluation of policies and procedures is applied on certain colleges to identify weaknesses and challenges of many methods and procedures of teaching. Educated approached can provide effective qualitative feedback to generate quantitative ratings of performance [45].

Romero *et al.* (2022) offered evaluation for educational process for different resources in the higher education. This research realizes that students are the main player in this evaluation process. The obtained results are analyzed in the period of 1 year. One of the big absorbed issues that applied different activities leading an effective impact for both students and teachers [46].

4. STATEMENT OF THE PROBLEM

The mission of the university is not only to teach and prepare learners but also to include research and community service, and to strive for its optimal development in the framework towards comprehensive development in various fields. Teaching is one of the most important functions of a faculty member. The evaluation of the performance of the faculty member is one of the issues that did not receive sufficient attention to researchers in the Arab countries compared to foreign studies. This is due to a fundamental reason in Arab societies, the faculty member considers it as a derogation.

The importance of the faculty member's role lies in the effective role of the faculty member in guiding students and enhancing their personal and cognitive development. Thus, the students must be given an important part in giving their opinion without restriction to stand on the reality of education and how to promote it. Therefore, the questionnaire is designed to reflect the opinions and concepts of students about the process of education and the faculty members.

5. METHODOLOGY

5.1. Questionnaire Design

A questionnaire was designed using typical format of five-level Likert test to obtain the views of the students in both the educational process and the faculty members. The questionnaire consists of four fields (Table 2):

- 1st field: Possibilities of the faculty member in teaching, ranking and preparing the material. This field including of six questions.
- 2nd field: Contribution of scientific material in the educational achievement of students. This field including of seven questions.
- 3rd field: Evaluation of the content of the scientific material. This field including of six questions.
- 4th field: Relationship between faculty member and students. This field including of six questions.

5.2. The Study Sample

Intentional sample was selected, that represented by the students of fourth stage of the Computer Science department and Information System department. These two departments belong to the Faculty of Computer and Information Technology at Anbar University. The fourth stage of both departments is selected as intentional sample because students at this stage reached the final stage and they are able to evaluate well in addition, only few days have passed since they received the university degrees.

This sample including two parts:

- Students at fourth sage of Computer Science department are 60 students divided into 24 (40%) males and 36 (60%) females
- Students at fourth sage of Information System department are 36 students divided into 15 (42%) males and 21 (58%) females.

6. RESULTS ANALYSIS AND DISCUSSION

The questionnaire is divided into four sets of questions as below:

- 1st set including questions (1–6) demonstrate the ability of faculty member in teaching
- 2nd set including questions (7–13) demonstrate the ability of the material in the achievement of students
- 3rd set including questions (14–19) evaluate the content of the scientific material
- 4th set including questions (20–25) demonstrate the relationship between faculty members and students.

The questionnaire is applied for all five subjects (each subject taught by separate teacher) of the fourth stage students in the department of computer science at university of Anbar.

6.1. Student responses analysis of Teaching Evaluation 1 (Image Processing Subject)

The histogram in Fig. 1a measures the ability of faculty member in teaching. This figure indicates that the overall weights are concentrated on the right side of the figure. That means most of the students answer in the agreement parts. In female section, it is clear that (36% agree and 46% strongly agree). In male section, it is clear that (41% agree and 52% strongly agree). The overall evaluation of this section gives (38% agree and 48% strongly agree). As a contribution, there is a slightly difference between the response of males and females.

The histogram in Fig. 1b measures the ability of the material in the achievement of students. This figure indicates that the overall weights are slightly shifted to the right side of the figure. That means a large number of the students answered in the agreement parts. In female section, it is clear that (46% agree and 21% strongly agree). In male section, it is clear that (55% agree and 32% strongly agree). The overall evaluation of this section gives (50% agree and 25% strongly agree). As a contribution there is a noticed difference between the response of males and females, in addition the response of males gives better agreement.

The histogram in Fig. 1c evaluates the content of the scientific material. This figure indicates that the overall weights are shifted to the right side of the figure. That means most of the students answered in the agreement parts. In female section it is clear that (47% agree and 39% strongly agree). In male section it is clear that (45% agree and 43% strongly agree). The overall evaluation of this section gives (47% agree and 40% strongly agree). As a contribution, there is a slightly difference between the response of males and females.

The histogram in Fig. 1d evaluates the relationship between faculty member and students. This figure indicates that the

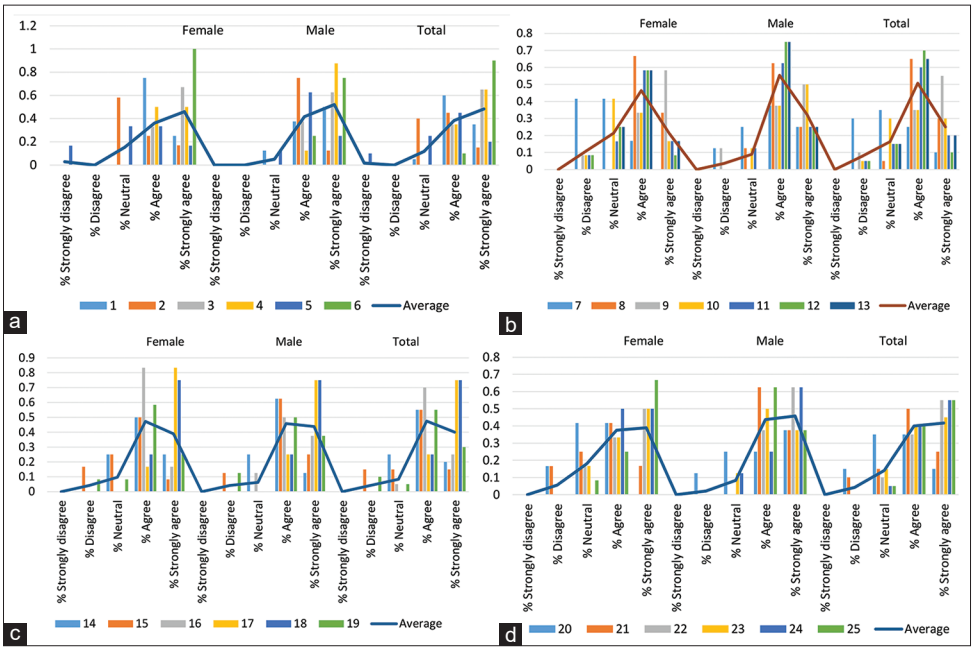


Fig. 1. Teaching Evaluation 1 computer science department. (a) Teaching Evaluation 1 (1st set), (b) Teaching Evaluation 1 (2nd set), (c) Teaching Evaluation 1 (3rd set), (d) Teaching Evaluation 1 (4th set).

overall weights are slightly shifted to the right side of the figure. That means big amount of the students answered in the agreement parts. In female section, it is clear that (38% agree and 39% strongly agree). In male section, it is clear that (44% agree and 46% strongly agree). The overall evaluation of this section gives (40% agree and 42% strongly agree). As a contribution, there is a slightly difference between the response of males and females. The response of males gives better result of agreement.

6.2. Student Responses Analysis of Teaching Evaluation 2 (Information Security Subject)

The histogram in Fig. 2a measures the ability of faculty member in teaching. This figure indicates that the overall weights are concentrated on the right side of the figure. That means most of the students answer in the agreement parts. In female section it is clear that (49% agree and 43% strongly agree). In male section, it is clear that (38% agree and 58% strongly agree). The overall evaluation of this section gives (41% agree and 48% strongly agree). As a contribution, there is a slightly difference between the response of males and females. In female, there is almost a normal distribution of answers between agree and strongly agree but in male there is a small orientation to strongly agree.

The histogram in Fig. 2b measures the ability of the material in the achievement of students. This figure indicates that the overall weights are concentrated on the center of the figure. That means a large number of the students answered in the neutral part. In female section, it is clear that (32% neutral, 36% agree, and 18% strongly agree). In male section, it is clear that (32% neutral, 28% agree, and 38% strongly agree). The overall evaluation of this section gives (32% neutral, 35% agree, and 26% strongly agree). As a contribution, there is a similarity between the response of males and females, except the strongly agree region have more voting with male students.

The histogram in Fig. 2c evaluates the content of the scientific material. This figure indicates that the overall weights are concentrated on the center of the figure. That means most of the students answered in the neutral region and both sides of agreement parts. In female section, it is clear that (18% disagree, 30% neutral, 18% agree, and 26% strongly agree). In male section, it is clear that (13% disagree, 40% neutral, 23% agree, and 25% strongly agree). The overall evaluation of this section gives (16% disagree, 34% neutral, 20% agree, and 26% strongly agree). As a contribution, there is almost a similarity between the response of males and females. In addition, there is a dissatisfaction of the content of the scientific material.

The histogram in Fig. 3d evaluates the relationship between faculty member and students. This figure indicates that the overall weights are slightly shifted to the right side of the

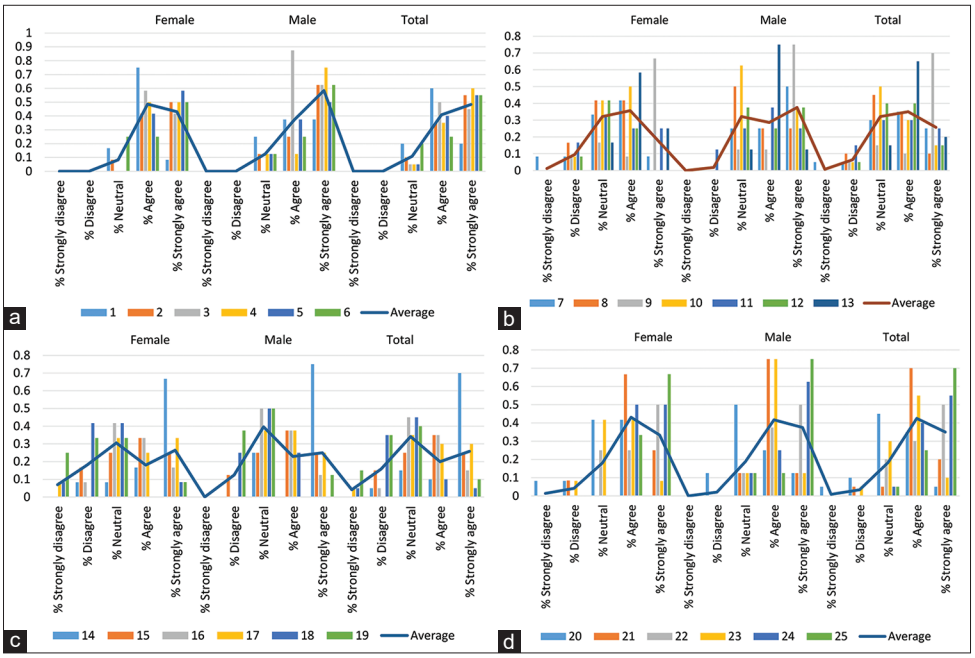


Fig. 2. Teaching Evaluation 2 computer science department, (a) Teaching Evaluation 2 (1st set), (b) Teaching Evaluation 2 (2nd set), (c) Teaching Evaluation 2 (3rd set), (d) Teaching Evaluation 2 (4th set).

figure. That means big amount of the students answered in the agreement parts. In female section, it is clear that (18% neutral, 43% agree, and 33% strongly agree). In male section, it is clear that (18% neutral, 42% agree, and 38% strongly agree). The overall evaluation of this section gives (18% neutral, 43% agree, and 35% strongly agree). As a contribution, there is a similarity in student response between males and females.

6.3. Student Responses Analysis of Teaching Evaluation 3 (ASP.net Subject)

The histogram in Fig. 3a measures the ability of faculty member in teaching. This figure indicates that the overall weights are slightly shifted on the right side of the figure. That means big amount of the students answer in the agreement parts. In female section, it is clear that (51% agree and 31% strongly agree). In male section, it is clear that (52% agree and 23% strongly agree). The overall evaluation of this section gives (53% agree and 27% strongly agree). As a contribution there is a slightly difference between the response of males and females, and the overall average curve is symmetry. In both female and male response values, there is almost a normal distribution of answers and most of the weights are oriented to agree part.

The histogram in Fig. 3b measures the ability of the material in the achievement of students. This figure indicates that the big amount of weights is concentrated on the center of the figure. That means there is significant number of the students answered in the neutral part. In female section, it is clear that (22% neutral, 41% agree and 32% strongly agree). In male section, it is clear that (25% neutral, 36% agree and 23% strongly agree). The overall evaluation of this section gives (23% neutral, 39% agree, and 23% strongly agree). As a contribution, there is a similarity between the response of males and females, except the strongly agree region have more voting with female students.

The histogram in Fig. 3c evaluates the content of the scientific material. This figure indicates that most of weights are concentrated on the center of the figure. That means big amount of students answered in the neutral region and both sides of agreement parts. In female section, it is clear that (10% disagree, 31% neutral, 32% agree, and 26% strongly agree). In male section, it is clear that (23% disagree, 21% neutral, 35% agree, and 10% strongly agree). The overall evaluation of this section gives (15% disagree, 27% neutral, 33% agree, and 20% strongly agree). As a contribution, there is no similarity between the response of males and females. In addition, there is a dissatisfaction of the content of the scientific material.

The histogram in Fig. 2d evaluates the relationship between faculty member and students. This figure indicates that the

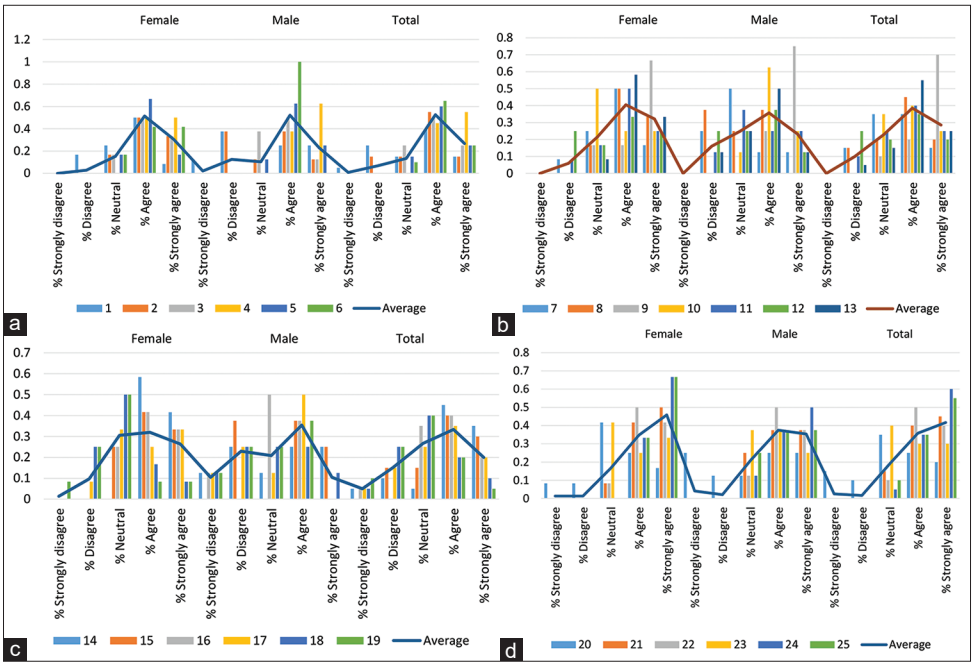


Fig. 3. Teaching Evaluation 3 computer science department, (a) Teaching Evaluation 3 (1st set), (b) Teaching Evaluation 3 (2nd set), (c) Teaching Evaluation 3 (3rd set), (d) Teaching Evaluation 3 (4th set).

overall weights are slightly shifted to the right side of the figure. That means big amount of the students answered in the agreement parts. In female section it is clear that (17% neutral, 36% agree, and 46% strongly agree). In male section it is clear that (21% neutral, 38% agree, and 36% strongly agree). The overall evaluation of this section gives (19% neutral, 36% agree, and 42% strongly agree). As a contribution, there is a similarity in student response between males and females with significant increases in female response.

6.4. Student Responses Analysis of Teaching Evaluation 4 (Operating Systems Subject)

The histogram in Fig. 4a measures the ability of faculty member in teaching. This figure indicates that the overall weights are concentrated on the center of the figure. That means big amount of the students answer in the agreement parts with a significant part of neutral. In female section it is clear that (20% neutral, 42% agree, and 22% strongly agree). In male section, it is clear that (29% neutral, 38% agree, and 31% strongly agree). The overall evaluation of this section gives (23% neutral, 40% agree, and 27% strongly agree). As a contribution, there is a slightly difference between the response of males and females, and the overall average curve is not very symmetry. In both female and male response values, there is almost a normal distribution of answers and most of the weights are oriented to agree part, in addition there is significant weight related to neutral part.

The histogram in Fig. 4b measures the ability of the material in the achievement of students. This figure indicates that the big amount of weights is concentrated on the center of the figure. That means there are significant number of the students answered in the neutral part, in addition there is a small weight related to disagree. In female section it is clear that (27% disagree, 14% neutral, 37% agree, and 12% strongly agree). In male section it is clear that (29% neutral, 45% agree, and 25% strongly agree). The overall evaluation of this section gives (17% disagree, 20% neutral, 40% agree, and 17% strongly agree). As a contribution, there is a dissimilarity between the response of males and females, and the average curve distributed almost over the figure.

The histogram in Fig. 4c evaluates the content of the scientific material. This figure indicates that most of weights are distributed on the span of the figure. That means big amount of students answered in all parts of the figure. In female section, it is clear that (15% disagree, 21% neutral, 28% agree, and 15% strongly agree). In male section, it is clear that (13% disagree, 38% neutral, 13% agree, and 31% strongly agree). The overall evaluation of this section gives (14% disagree, 28% neutral, 22% agree, and 22% strongly agree). As a contribution, there is no similarity between the response of males and females. In addition, there is a dissatisfaction of the content of the scientific material and there is a significant weight related to neutral, disagree, and strongly disagree parts.

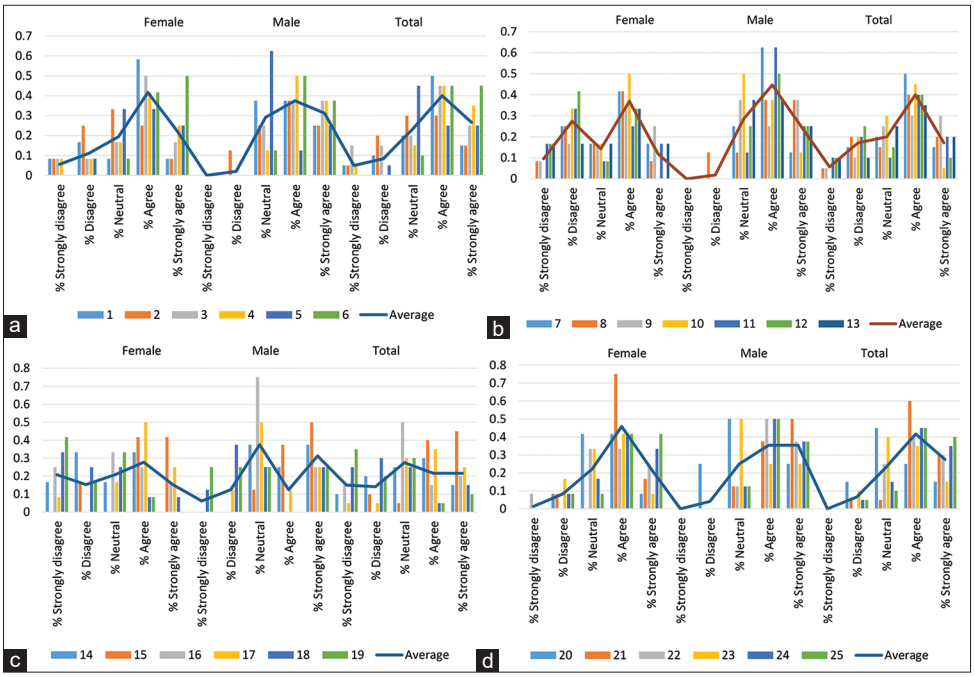


Fig. 4. Teaching Evaluation 4 computer science department, (a) Teaching Evaluation 4 (1st set), (b) Teaching Evaluation 4 (2nd set), (c) Teaching Evaluation 4 (3rd set), (d) Teaching Evaluation 4 (4th set).

The histogram in Fig. 4d evaluates the relationship between faculty member and students. This figure indicates that the overall weights are slightly shifted to the right side of the figure. That means big amount of the students answered in the agreement parts. In female section, it is clear that (22% neutral, 46% agree, and 22% strongly agree). In male section, it is clear that (25% neutral, 35% agree, and 35% strongly agree). The overall evaluation of this section gives (23% neutral, 42% agree, and 28% strongly agree). As a contribution, there is a similarity in student response between males and females with significant increases in male response. In addition, there is significant weight in the neutral part.

6.5. Student Responses Analysis of Teaching Evaluation 5 (Artificial Intelligence Subject)

The histogram in Fig. 5a measures the ability of faculty member in teaching. This figure indicates that the overall weights are slightly shifted on the right side of the figure. That means big amount of the students answer in the agreement parts. In female section, it is clear that (10% neutral, 35% agree, and 51% strongly agree). In male section, it is clear that (27% neutral, 50% agree, and 19% strongly agree). The overall evaluation of this section gives (17% neutral, 40% agree, and 39% strongly agree). As a contribution, there is a slightly difference between the response of males and females, and the overall average curve is symmetry. In both female and male response values, there is almost a normal distribution of answers and most of the weights are oriented to agree part, in addition there is significant weight related to the neutral part exactly in male response.

The histogram in Fig. 5b measures the ability of the material in the achievement of students. This figure indicates that the big amount of weights is concentrated on the right side of the figure. That means most of the student answered on the agreement part in addition there are significant number of the students answered in the neutral part. In female section, it is clear that (14% neutral, 50% agree, and 35% strongly agree). In male section, it is clear that (30% neutral, 43% agree, and 16% strongly agree). The overall evaluation of this section gives (21% neutral, 47% agree, and 27% strongly agree). As a contribution, there is a dissimilarity between the response of males and females, in addition there is a significant weight related to neutral part exactly related to male response.

The histogram in Fig. 5c evaluates the content of the scientific material. This figure indicates that most of weights are concentrated on the center of the figure. That means big amount of students answered in the neutral region and both sides of agreement parts. In female section, it is clear that (11% disagree, 25% neutral, 44% agree, and 17% strongly agree). In male section, it is clear that (31% disagree, 38%

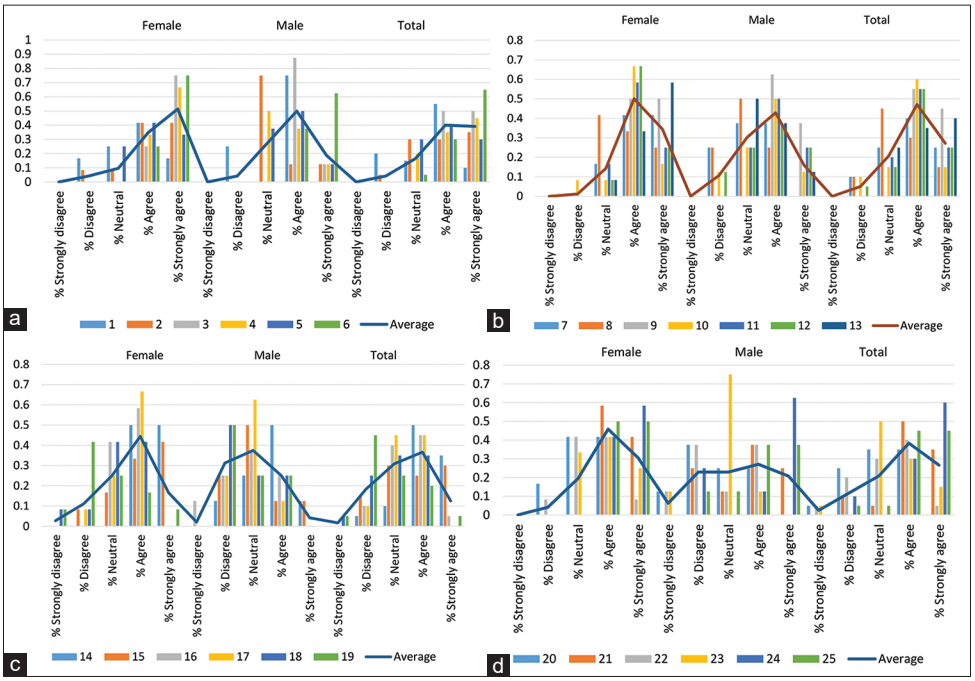


Fig. 5. Teaching Evaluation 5 computer science department, (a) Teaching Evaluation 5 (1st set), (b) Teaching Evaluation 5 (2nd set), (c) Teaching Evaluation 5 (3rd set), (d) Teaching Evaluation 5 (4th set).

neutral, 25% agree, and 42% strongly agree). The overall evaluation of this section gives (18% disagree, 31% neutral, 37% agree, and 13% strongly agree). As a contribution, there is no similarity between the response of males and females. In addition, there is a dissatisfaction of the content of the scientific material.

The histogram in Fig. 5d evaluates the relationship between faculty member and students. This figure indicates that the overall weights are concentrated on the center of the figure. That means the students answered are distributed on all parts of the figure. In female section, it is clear that (4% disagree, 19% neutral, 46% agree, and 31% strongly agree). In male section, it is clear that (23% disagree, 23% neutral, 27% agree, and 21% strongly agree). The overall evaluation of this section gives (2% disagree, 21% neutral, 38% agree, and 27% strongly agree). As a contribution, there is no similarity in student response between males and females with significant increases of disagree in male response that explains a weak relationship between faculty member and students.

7. CONCLUSIONS

Assessing the learning process is an important topic that always occupies the educational administration. This process depends on a comprehensive re-evaluation of the courses, teachers, and department management. This evaluation applied on the fourth stage (final stage) students of computer science department. All subjects of fourth stage are taught by high skilled Ph.D. teachers, in addition these subjects are well organized and updated.

The obtained results realized that the fourth stage students have the ability to take their decisions in a right way. In addition, that the most received responses (about 80%) are concentrated on agree and strongly agree, but with some reservation of the teaching evaluation 5 that their answers were distributed almost equally between neutral, agree and strongly agree with small weight to disagree. By revealing the reason for this response to teaching evaluation 5, we can conclude that this result was obtained due to inexperience of teachers. However, this is the 1st time at this department implemented an evaluation guided by students to reflect their opinion about the teaching process. This process can be circulated to include all departments and colleges at the university.

REFERENCES

[1] K. Zamoum. "Teaching crisis management in Arab universities: A critical assessment". *Public Relations Review*, vol. 39, pp. 47-54, 2013.

[2] T. Kabakian-Khasholian, A. El-Nemer and H. Bashour. "Perceptions

about labor companionship at public teaching hospitals in three Arab countries". *International Journal of Gynecology and Obstetrics*, vol. 129, pp. 223-226, 2015.

[3] A. Addi-Raccah and Y. Grinshtain. "Teachers' capital and relations with parents: A comparison between Israeli Jewish and Arab teachers". *Teaching and Teacher Education*, vol. 60, pp. 44-53, 2016.

[4] A. Al-Krenawi. "Higher education among minorities: The Arab case". *International Journal of Educational Research*, vol. 76, pp. 141-146, 2016.

[5] S. F. A. Bakhiet, E. Dutton, K. Y. A. Ashaer, Y. A. S. Essa and G. Madison. "Understanding the simber effect: Why is the age-dependent increase in children's cognitive ability smaller in Arab countries than in Britain"? *Personality and Individual Differences*, vol. 122, pp. 38-42, 2018.

[6] H. Khalidi and B. A. S. Dajani. "Facets from the translation movement in classic Arab culture". *Procedia Social and Behavioral Sciences*, vol. 205, pp. 569-576, 2015.

[7] Jingning, Z. "Promotion criteria, faculty experiences and perceptions: A qualitative study at a key university in China". *International Journal of Educational Development*, vol. 33, pp. 185-195, 2013.

[8] Y. Feniger and H. Ayalon. "English as a gatekeeper: Inequality between Jews and Arabs in access to higher education in Israel". *International Journal of Educational Research*, vol. 76, pp. 104-111, 2016.

[9] M. Engin and S. Donanci. "Dialogic teaching and iPads in the EAP classroom". *Computers and Education*, vol. 88, pp. 268-279, 2015.

[10] M. Dickson, H. Kadbey and M. McMinn. "Comparing reported classroom practice in public and private schools in the United Arab Emirates". *Procedia Social and Behavioral Sciences*, vol. 186, pp. 209-215, 2015.

[11] M. T. Abuelma'atti. "Evaluation of engineering research in Arab countries using a bibliometric-based approach". *Procedia Social and Behavioral Sciences*, vol. 102, pp. 438-445, 2013.

[12] N. Arsad, N. Kamal, N. Saibani, A. Ayob and H. Husain. "Student performance early indicator using mathematics and engineering fundamental test". *Procedia Social and Behavioral Sciences*, vol. 102, pp. 86-91, 2013.

[13] T. R. Dikheel and H. S. Uraibi. "Effects of fundamental schools on mathematics performance: Iraq as a model". *Procedia Social and Behavioral Sciences*, vol. 8, pp. 236-241, 2010.

[14] H. Abood and M. S. Al-Ani. "Obstacles to the spread of e-learning in the Arab countries". *Journal of Education and Learning*, vol. 10, pp. 347-354, 2016.

[15] E. Van Ommering. "Teaching on the frontline: The confines of teachers' contributions to conflict transformation in Lebanon". *Teaching and Teacher Education*, vol. 67, pp. 104-113, 2017.

[16] I. Abu-Saad. "Access to higher education and its socio-economic impact among Bedouin Arabs in Southern Israel". *International Journal of Educational Research*, vol. 76, pp. 96-103, 2016.

[17] A. M. Soliman. "Appropriate teaching and learning strategies for the architectural design process in pedagogic design studios". *Frontiers of Architectural Research*, vol. 6, pp. 204-217, 2017.

[18] M. M. Awad, W. S. Salem, M. Almuhaizaa and Z. Aljeaidi. "Contemporary teaching of direct posterior composite restorations in Saudi dental schools". *The Saudi Journal for Dental Research*, vol. 8, pp. 42-51, 2017.

[19] S. Khouyibaba. "Teaching remedial courses: Challenges and teaching philosophy". *Procedia Social and Behavioral Sciences*, vol. 186, pp. 927-931, 2015.

[20] S. E. Cristache, D. Serban and M. Vuta. "An analysis of the Romanian high education system perspectives using quantitative techniques". *Procedia Social and Behavioral Sciences*, vol. 186, pp. 53-57, 2015.

[21] Z. Zhang and Y. Xue. "An investigation of how Chinese university students use social software for learning purposes". *Procedia Social and Behavioral Sciences*, vol. 186, pp. 70-78, 2015.

[22] A. Kokkos. "The challenges of adult education in the modern world". *Procedia Social and Behavioral Sciences*, vol. 180, pp. 19-24, 2015.

[23] N. I. Jabbouri, R. Siron, I. Zahari and M. Khalid. "Impact of information technology infrastructure on innovation performance: An empirical study on private universities in Iraq". *Procedia Economics and Finance*, vol. 39, pp. 861-869, 2016.

[24] S. A. David, H. Taleb, S. S. A. Scatolini, A. J. Al-Qallaf, M. A. George and H. S. Al-Shammari. "An exploration into student learning mobility in higher education among the Arabian gulf cooperation council countries". *International Journal of Educational Development*, vol. 55, pp. 41-48, 2017.

[25] F. Suja, Z. Yacob and A. Mohammed. "Background factors contributing to performance of master course students-case study of civil engineering master programme". *Procedia Social and Behavioral Sciences*, vol. 60, pp. 382-389, 2012.

[26] R. Seginer and S. Mahajna. "On the meaning of higher education for transition to modernity youth: Lessons from future orientation research of Muslim girls in Israel". *International Journal of Educational Research*, vol. 76, pp. 112-119, 2016.

[27] D. Young and S. Seibenhener. "Preferred teaching strategies for students in an associate of science nursing program". *Teaching and Learning in Nursing*, vol. 13, pp. 41-45, 2018.

[28] S. A. Lannan. "Nursing program evaluation for nurse educators". *Nurse Education Today*, vol. 55, pp. 17-19, 2017.

[29] C. Lawson, S. Pati, J. Green, G. Messina, A. Strömberg, N. Nante, D. Golinelli, A. Verzuri, S. White, T. Jaarsma, P. Walsh, P. Lonsdale and U. T. Kadam. "Development of an international comorbidity education framework". *Nurse Education Today*, vol. 55, pp. 82-89, 2017.

[30] F. Naja, H. Shatila, L. Meho, M. Alameddine, N. Hwalla, S. Haber, L. Nasreddine and A. M. Sibai. "Gaps and opportunities for nutrition research in relation to non-communicable diseases in Arab countries: Call for an informed research agenda". *Nutrition Research*, vol. 47, pp. 1-12, 2017.

[31] A. Sari, A. Firat, A. Karaduman. "Quality assurance issues in higher education sectors of developing countries; Case of Northern Cyprus". *Procedia Social and Behavioral Sciences*, vol. 229, pp. 326-334, 2016.

[32] S. Liu and J. Liu. "Quality assurance in Chinese higher education". In: *The Rise of Quality Assurance in Asian Higher Education*. Ch. 2. Elsevier, Amsterdam, Netherlands, pp. 15-33, 2017.

[33] G. Menon. "Maintaining quality of education in management institutes-reforms required". *Procedia Social and Behavioral Sciences*, vol. 133, pp. 122-129, 2014.

[34] M. C. Pana and C. Mosora. "From quantity to quality in addressing the relationship between education and economic development". *Procedia Social and Behavioral Sciences*, vol. 93, pp. 911-915, 2013.

[35] C. Y. Fook and G. K. Sidhu. "Learning practices in a higher learning institute in United States". *Procedia Social and Behavioral Sciences*, vol. 90, pp. 88-97, 2013.

[36] A. A. M. Allah aalYateem and N. B. B. Hameed. “Digital repositories in the Arab universities: A comparative analytical study”. *Procedia Computer Science*, vol. 65, pp. 768-777, 2015.

[37] M. E. Kiersma, A. M. H. Chen, E. L. Kleppinger, E. W. Blake, N. M. Fusco, V. Mody, M. E. Gillespie, M. Knell and R. M. Zavod. “Evaluation of criteria utilized in the recognition of teaching excellence awards”. *Currents in Pharmacy Teaching and Learning*, vol. 8, pp. 477-484, 2016.

[38] A. Barana and M. Marchisio. Ten good reasons to adopt an automated formative assessment model for learning and teaching mathematics and scientific disciplines”. *Procedia Social and Behavioral Sciences*, vol. 228, pp. 608-613, 2016.

[39] J. Artés, F. Pedraja-Chaparro and M. del Mar Salinas-JiméneZ. “Research performance and teaching quality in the Spanish higher education system: Evidence from a medium-sized university”. *Research Policy*, vol. 46, pp. 19-29, 2017.

[40] F. N. A. Alhija. “Teaching in higher education: Good teaching through students’ lens”. *Studies in Educational Evaluation*, vol. 54, pp. 4-12, 2017.

[41] L. Wikander and S. L. Bouchoucha. “Facilitating peer based learning through summative assessment-an adaptation of the objective structured clinical assessment tool for the blended learning environment”. *Nurse Education in Practice*, vol. 28, pp. 40-45, 2018.

[42] X. N. C. Nguyen and I. Walkinshaw. “Autonomy in teaching practice: Insights from Vietnamese English language teachers trained in inner-circle countries”. *Teaching and Teacher Education*, vol. 69, pp. 21-32, 2018.

[43] J. D. Cano-Moreno, J. M. Arenas, V. Sánchez, M. Islán, J. Narbón. “Methodology for quantitative evaluation of university teaching. Application to the subject of project management”. *Procedia Manufacturing*, vol. 41, pp. 930-937, 2019.

[44] G. Li, L. Xiang, Z. Yu and H. Li. “Intelligent evaluation of teaching based on multi-networks integration”. *International Journal of Cognitive Computing in Engineering*, vol. 1, pp. 9-17, 2020.

[45] L. Lohman. “Evaluation of university teaching as sound performance appraisal”. *Studies in Educational Evaluation*, vol. 70, pp. 1-11, 2021.

[46] S. Romero, G. Camargo, K. Garcia, C. Medina, I. Hernande and J. Grimaldo. “Analysis of the students’s appreciation of teaching as a tool for managing educational processes”. *Procedia Computer Science*, vol. 203, pp. 537-543, 2022.

SURVEY

UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Log File Analysis Based on Machine Learning: A Survey

Rawand Raouf Abdalla, Alaa Khalil Jumaa

Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq

ABSTRACT

In the past few years, software monitoring and log analysis become very interesting topics because it supports developers during software developing, identify problems with software systems and solving some of security issues. A log file is a computer-generated data file which provides information on use patterns, activities, and processes occurring within an operating system, application, server, or other devices. The traditional manual log inspection and analysis became impractical and almost impossible due logs' nature as unstructured, to address this challenge, Machine Learning (ML) is regarded as a reliable solution to analyze log files automatically. This survey tries to explore the existing ML approaches and techniques which are utilized in analyzing log file types. It retrieves and presents the existing relevant studies from different scholar databases, then delivers a detailed comparison among them. It also thoroughly reviews utilized ML techniques in inspecting log files and defines the existing challenges and obstacles for this domain that requires further improvements.

Index Terms: Log Files, Log Analysis, Machine Learning, Anomaly Detection, User Behavior, Log File Maintenance

1. INTRODUCTION

In the context of computing, logs are bits of data that give insight into numerous events that occur during the execution of a computer program [1]. Information technology utilization has increased at an unparalleled rate during the previous two decades. Data of various types are shared through a broad range of networks, from company-wide LAN networks to public hub wireless access networks. As the transmission and consumption of data through these networks grows, so does number of breaches and network intrusion efforts aimed at obtaining secret and personal information. As a consequence of this, security for networks and data has become a highly significant topic in both the academic and practical computing communities [2].

Log data comprised more than 1.4 billion logs each day is used to detect suspicious business-specific activities and user profile behavior [3].

A series of devices and software generate log files in dissimilar formats. Log files are used by software systems to retain track of their activities. Different system part, like OS, may record its events to a remote log server. An OS is the machine software that controls computer hardware and software resources and permits the execution of multiple applications [4]. The start or end of occurrences or activities of software system, status information, and error information are all captured in the log files. User information, application information, date and time information, and event information are normally included in each log line. When these files are properly analyzed, they may provide important information about numerous characteristics every system. For monitoring, troubleshooting, and problem detection, logs are often gathered [5].

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.77-84

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: rawand.raouf.a@spu.edu.iq

Received: 16-07-2022 Accepted: 07-09-2022 Published: 07-10-2022

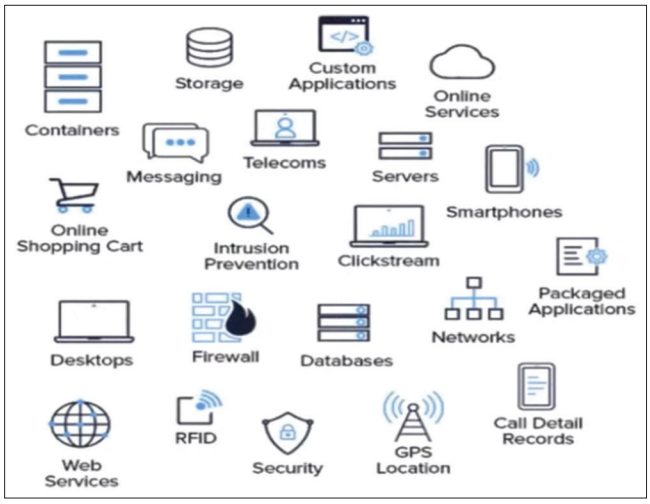


Fig. 1. Some of log file sources [11].

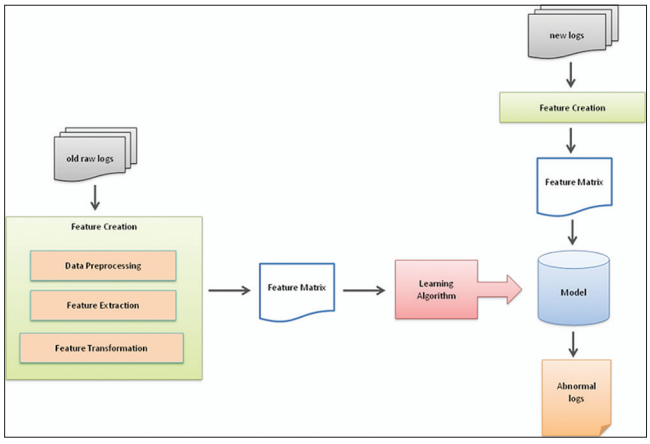


Fig. 2. Overview of the learning system [14].

Normally, log files are saved as text, compressed, or binary files. The most commonly used format is text files, which have the gains of utilizing fewer CPU and I/O resources when producing files, allowing for long-term storage and maintenance, and being easy to read and use. The binary format is a machine-readable log file format created by a platform that requires a particular tool to view, making it unsuitable for long-term team storage. While compressing log files, use an appropriate compression format standard with multi-platform compatibility for efficient log storage and usage [6].

Log files record activity on different kinds of servers, including data servers and application servers. Because log files record a range of server actions and include a huge data in the form of server-produced messages, they are often relatively big files. These messages include valuable knowledge, like what apps were operating on the server, when they were run, and by whom a log file contains a lot of messages that indicate various server actions. In addition, each unique message type will have hundreds, if not thousands, of entries in the log file, each with slight variances in the general structure. The system administrator may utilize these messages for a lot of reasons, such as intrusion detection, reporting, performance monitoring, anomaly detection, and so on.

2. STATE OF THE ART

This survey tries to explore the most recent existing studies on analysis log files by using both of supervised and unsupervised machine learning algorithms. Different models and techniques have been proposed by researchers to different aspects. Several studies have utilized techniques and models to predict attack, user behavior, and system failure to increase server security and systems, marketing, and decrease failure time. This study revealed that there are many gaps which require further improvements such as: Using real dataset in creating models, more log analysis, or mining must be done to obtain meaningful information and minimize the false positive and negative results and the maintenance aspect requires further improvements compared to the other mentioned aspects.

3. COMMON LOG FILE TYPES

Almost every network device produces a unique form of data, and each component logs those data in its own log. As a result, there are several types of logs, such as [7]:

3.1. Application Logs

Developers have a strong grip on application logs. It may include any kind of event, error message, or warning that the program generates. Application logs provide information to system administrators concerning the status of an application running on a server. Application logs should be well-structured, event-driven, and include pertinent data to assist as the center for higher-level abstraction, visualization, and aggregation. The application logs' event stream is required for viewing and filtering data from numerous instances the programs.

3.2. Web Server Logs

Every user communication with the web is saved as a record in a log file called a "web log file" that is a text file with the extension ".txt." The data created automatically of users' interactions with the web, and will be saved in a variance log files, including server access logs, error logs, referrer logs, and client-side cookies. In the style of text file, this web log

TABLE 1: A list of studies for the purpose of (identifying user behavior)

Reference No.	Classification Algorithms	Performance
[16]	NN	Prediction Accuracy is 90%
[17]	(Parzen) (Gauss) (PCA) and (KMC)	Prediction Accuracy for daily activity dataset is 90% e-mail content dataset is 65% e-mail communication network dataset 75%
[18]	Modified Span Algorithm and the Personalization Algorithm	Provides high predication accuracy

TABLE 2: A list of studies for the purpose of (Security)

Reference No.	Classification Algorithms	Performance
[20]	K-Means Clustering	Direction Accuracy ratio is 83%
[21]	SVR, LR, and KNR	Offers excellent security protection
[22]	LR, NN, RF, and XG	Direction Accuracy ratio is 85% with 0.78% false positive rate
[23]	SVM	Direction Accuracy ratio is 99%
[24]	K-Means Clustering	Direction Accuracy ratio for SOM#34 is (84.37%) AAU is (90.01%)

TABLE 3: A list of studies for the purpose of (Maintenance)

Reference No.	Classification Algorithms	Performance
[25]	Discovering Patterns from Temporal Sequences	Provides high system performance
[26]	Principal Component Analysis (PCA)	Provides high system performance

saves all and every web request executed by the client to the servers. Each line or record in the web log file links to a user's request to the servers. The logs file for the web data are: Web Server Logs, Proxy Server Logs, and Browser Logs [8].

Web server logs typically include the IP address, the date and time of request, the exact request line providing by users, the URL, and the requested file type.

3.3. System Logs

The OS records specified events in System log. In addition, these logs are an excellent resource for obtaining information about external events. Typically, a system log includes entries generated by the OS, such as system failures, warnings, and errors. Individual programs may generate log files related with user sessions that include data on the user's login time, interactions with the application, authentication result, and so on. While an operating system-generated log file is mentioned to as a system log, files produced by particular programs or users are related to as audit data. Examples include records of successful and unsuccessful login attempts, system calls, and user command executions [9].

3.4. Security Logs

Security logs are utilized to give enough capabilities for identifying harmful actions after their occurrence with the intention of prevent them from recurrence. Security logs preserve track of a range information that has been pre-defined by system administrators. For example, firewall logs contain information about packets routed from their sources, rejected IP addresses, outbound activity from internal systems, and failed logins. Security logs contain detailed information that security administrators must manage, regulate, and evaluate in conformity with their requirements [7].

3.5. Network Logs

Network Logs offer different information on various events that occurred on the networks. Among the events are the recording of malicious activity, a rise in network traffic, packet losses, and bandwidth delays. Network logs may be gathered from a range of network devices, including switches, routers, and firewalls. By monitoring network logs for various attack attempts, network administrators can monitor and troubleshoot normal networking.

3.6. Audit Logs

Record any network or system activity that is not allowed in a sequential order. It aids security managers in analyzing malicious activity during an attack. The source and destination addresses, timestamp and user login information are usually the most significant parts of information in audit log files.

3.7. Virtual Machine Logs

This log files include details on the instances that are executing on the VM, such as their startup configuration, operations, also the time they complete their execution. The VM logs retain track of many processes, including the number of instances operating on the VM, the execution duration of each application, and application migration, which assists the CSP in identifying malicious activity that happened while the attack [10]. Figure 1 show some of log file sources.

4. LOG MINING

Log mining is a technique which employs statistics, data mining, and ML to automatically explore and analyze vast amounts of log data in essence to find useful patterns and trends. The data and tendencies gleaned might assist in the monitoring, administration, and troubleshooting of software systems [12]. Web Usage Mining (WUM) as an example because it the most frequently utilized log file types.

Web mining is separated into three categories: Web Usage Mining, Web Content Mining and Web Structure Mining. Web usage mining is a procedure for capturing web page access data. The pathways leading to viewed web sites are providing by this use data. This data are often collected automatically by the web server then stored in access logs. Other important information provided by Common Gateway Interface CGI scripts includes referral logs, survey log data, and user subscription information. This area is significant to the entire usage of data mining by businesses, institutions, and their data access and web-based applications. Three steps necessity be taken in Web Usage Mining which are [13]:

4.1. Data Preprocessing

The web logs contain raw data that cannot be utilized to generate information. During this step, engineers use techniques to transform original data for a usable format. Typically, real-world data are incomplete, unexpected, and lacking of behavior or patterns, in addition including many mistakes. Data pre-processing is a tried-and-true way of solving these issues.

4.2. Pattern Discovery

Those pre-processing results are then used to determine a pattern of frequent user access. To identify significant information, several data mining methods for instance association rules, clustering, classification, and sequential pattern approach will be used in pattern discovery. The information that has been obtained could be presented in a number of ways including graphs, charts, tables, and so on.

4.3. Pattern Analysis

Final outcome of the pattern discovery step are not utilized directly in the analysis. Accordingly, during this phase, a strategy or tool will be developed to assist analysts in comprehending the knowledge which has been gathered. Visualization approaches, Online Analytical Processing OLAP analysis, and this phase might involve the use of tools or methods for instance knowledge query mechanisms. Figure 2 Overview of the learning system.

5. TYPES OF LOG FORMAT

Common servers use one of these three types of log file formats [15]:

5.1. Common Log File Format

Web servers create log files utilizing this standardized text file format. The setup of the standard log file format is provided in the box that follows.

Example of Common Log File Format [15].

5.2. Combined Log Format

It is like the previous log file format with the add of the referral field, user agent field, and cookie field. The setup for this format is shown in the box follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{Useragent}i\"" combined CustomLog log/access_log combined eg: 127.0.0.1 - frank [15/Oct/2021:14:59:38 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2328 "https://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
```

Example of Combined Log Format [15].

5.3. Multiple Access Logs

It is a hybrid of the common log and the combined log file format, with the ability to establish several directories for access logs. The structure of various access logs is detailed in the box that follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common CustomLog logs/referer_log "%{Referer}i -> %U" CustomLog logs/agent_log "%{User-agent}i"
```

Example of Multiple Access Logs [15].

6. DATA PREPROCESSING AND ML

6.1. Data Preprocessing

It is critical to preprocess data to handle with different flaws in raw gathered data, which may include noise such as mistakes, redundancies, outliers, and other missing values or unclear data. The most prevalent procedures in data preprocessing are [16]:

6.1.1. Data cleaning
Handle data inconsistencies, noise, and missing values.

6.1.2. Data integration
Seeks to integrate data from several sources into a cohesive data storage unit. Which is not an easy operation, since it entails establishing compatibility across several schema types. Weak or inefficient data integration might result in inconsistency and redundancy, while a well-implemented solution would surely improve accuracy and improve subsequent operations. Data integration techniques involve entity identification, correlation analysis, tuple deduplication, redundancy, and along with the discovery and resolution of data value conflicts.

6.1.3. Data transformation
Aims to transform data in a style which is both useable and meaningful format. The reason is for data mining processes to be more efficient. Smoothing, feature building, normalization, discretization, and generalization of nominal data are all examples of data transformation strategies. These subtasks are heavily reliant on the preprocessed data and need human supervision.

6.2. Machine Learning
A science focused with the theory, performance, and features of learning systems and algorithms. ML is a highly interdisciplinary field that depend on techniques from several fields, including artificial intelligence, cognitive science, optimization theory, information theory, statistics, and optimal control. ML has permeated virtually every scientific subject on consequence of its broad use in a various of applications, having a tremendous impact on both research and society. It was used to a range of challenges, such as autonomous control systems, informatics and data mining, recognition systems, and recommendation systems [17].

ML is broadly classified into three subfields [18]:
- Supervised Learning: It needs training on labeled data that contain both inputs and outputs.
- Unsupervised Learning: Not needs labeled training data, as the environment solely offers unlabeled inputs.
- Reinforcement Learning: It permits learning to occur on consequence of feedback obtained from interactions with the external environment.

Analysis of log files relevant to a failed execution may be laborious, particularly if the file contains thousands of lines. Utilizing current advancements in text analysis using deep neural networks (DNN), research [3] presents an approach to decrease the effort required to study the log file by highlighting the most likely informative content in the failed log file, which may aid in troubleshooting the failure's causes. In essence, they decrease the size of the log file by deleting lines deemed to be of less significance to the problem.

7. CONTRIBUTION OF THE LOG ANALYSIS

The log analysis contribution split into four categories, as follows [19]:

7.1. Performance
Used to find the system's performance during the optimization or troubleshooting phase. In the instance of performance, logs assist the administrator in clarifying how a specific system's resource has been utilized.

7.2. Security
Security logs are a lot used to detect security breaches or misconduct and to conduct postmortem investigations into security occurrences. For example, intrusion detection requires reconstructing sessions from logs that identify illegal system access.

7.3. Prediction
In addition, logs have ability of producing predictive information. There are predictive analytic systems that utilize log data to assist with marketing plan development, advertising placement, and inventory management.

7.4. Reporting and Profiling
Furthermore, log analysis is required for analyzing resource usage, workload, and user activity. For instance, logs will capture the attributes of jobs inside a cluster's workloads to profile resources utilize within large data center.

8. SURVEY METHODOLOGY

Collect articles for this research have been done in a systematic manner comprehensive database for the research on automated log analysis was utilized. Relevant articles were identified in online digital libraries, and the repository was extended manually by evaluating the references to these articles. The libraries can now be accessed online. To begin, looked through a range of well-known online digital repositories (e.g., ACM Digital Library, Elsevier Online, ScienceDirect, IEEE Xplore, Springer Online, and Wiley Online). According to these studies, most prevalent uses of log files with ML algorithms is classified into numerous categories, on which we based our study:

8.1. Identify User Behavior

User activity analysis using logs may provide significant information about users. User clustering based on logs enables the gathering of clients considering their activity and subsequent analysis of user access patterns, making it an excellent option for problem solving [20].

Xu *et al.* [21] examine use of HTTP traffic to find the identities of users. Techniques presuppose access to a proxy server's log. Thus, it is likely to develop web use profiles for people who utilize devices with a static IP address. They demonstrated that given a web use profile, it is feasible to identify users on any other device or to monitor when another user uses a device. Technically, they divide web traffic across sessions that link to the traffic of a distinct IP address over a definite time period. They reduce every session to a frequency vector distributed over the vector space of accessible domain. They used a set of methods for instance-based user identification centered on this representation. Experiments showed that centered on gathered web usage profiles using Nearest Neighbor classification, user identification is achievable with a prediction accuracy of greater than 90%. This paper needs to examine the usage of more sophisticated identification and obfuscation methods integrating the time series of URLs more closely.

Kim *et al.*[22], based on user behavior modeling and anomaly detection techniques, the authors offered a framework for detecting insider threats throughout the user behavior modeling process. They constructed three datasets depending on the CERT database: Users daily action dataset, an e-mail content dataset, and an e-mail communication dataset depending on the user account and sending and receiving information. They proposed insider-threat identification models using those datasets, applying ML set anomaly detection methods to imitate real-world companies with just a few potentially harmful insiders' activities. In this work, the authors employed classification algorithms for insider-threat detection. The findings in this study recommend that the suggested framework is capable of detecting malicious insider behaviors relatively effectively. On the basis of the daily activity summaries dataset, the anomaly detection achieved a maximum detection 90% percent by monitoring top 30% of anomaly. According to the e-mail content datasets, the detection 65.64 % detected while 30% of sceptical e-mails have been monitored. The paper's limitation is that, although the dataset (CERT) used to building the system was carefully developed and contains a variety of threat scenarios, it stills an artificially and simulated produced dataset.

Prakash *et al.* [23] investigated for the scope of analyzing user prediction behavior based on users personalization obtained from web logs. A web log records the user's navigation patterns when visiting websites. The user navigation pattern could be analyzed using the user's recent weblog navigation. The weblog has several posts with data such as the status code, IP address, and amount of bytes sent, along with categories and a time stamp. User interests could be categorized according to categories and attributes, which aids in determining user behavior. The goal of this research is to differentiate between interested and uninterested user behavior through classification. The Modified Span Algorithm and the Personalization Algorithm are used to identify the user's interest. Table 1 provides a summary list of studies we reviewed for the purpose of identifying user behavior.

8.2. Security Issues

Recently, some researchers and programmers utilizing data mining methods to log-based Intrusion Detection Systems (IDS) resulted in a powerful anomaly detection-based (IDS) which depended solely on the inflowing stream of logs to discern what may be normal and what is not (possibly an attack) [24].

Zeufack *et al.* [25] offered a fully unsupervised framework for real-time detection of abnormalities. This concept is separated into two phases: A knowledge base development stage, that use clustering to identify common patterns, and A streaming anomaly detection stage that detects abnormal occurrences in real time. They test their framework on (Hadoop Distributed File System) log files and it successfully detects anomalies with an F-1 score of 83%. This framework ought to be improved to get advantages for other features that are embedding in a log file and has positive impact on anomalies detection.

The authors of [26] presented a Dempster-Shafer (D-S) evidence theory-based host security analysis technique. They acquire information of monitoring logs and use it to design security analysis model. They utilize three regression models as sensors for multi-source information fusion: Logistic regression, support vector regression, and K-nearest neighbor regression. The suggested technique offers excellent strong security for host. Improved ML approaches may increase accuracy of evidence in this research, resulting in more accurate probability values for host security analysis.

Study [27] a ML-based system for identifying insider threats in organizations' networked systems is provided. The research discussed four ML algorithms: Neural Networks (NN), Random Forest (RF), Logistic Regression (LR), and

XGBoost (XG) across multiple data granularities, limited ground truth, and training scenarios to assist cyber security analysts in detecting malicious insider behaviors in unseen data. Evaluation results showed that the proposed system can successfully learning from the limited training data and generalize to detect new users with malicious behaviors. The system has a great detection rate and precision, mainly when user-generated findings are considered. The downside: Will examine the utilization of temporal information in user activities. Specifically, all the systems in this research gave labels based on a single exemplar's state description. Allowing models to view many exemplars or to maintain state (recurrent connections) can allow models to make non-Markovian decisions.

Shah *et al.* [28] offered an expanded risk management strategy for Bring Your Own Device (BYOD) to increase the safety of the device environment. The proposed system makes usage mobile device management system, system logs, and risk management systems to detect malicious activities using machine learning. They can state that the result achieved 99% detection rate with the practice of Support Vector Machine algorithm.

Tadesse *et al.* [29] employed multilayer log analysis to discover assaults at several stages of the datacenter. Thus, identifying distinct assaults requires considering the heterogeneity of log entries as an initial point for analysis. The logs were integrated in a common format and examined based on characteristics. Clustering and Correlation are the root of the log analyzer in the center engine, which operate alongside the attack knowledge base to detect attacks. To calculate the quantity of clusters and filter events according on the filtering threshold, clustering methods for instance Expectation Maximization and K-means were utilized. On the furthermore, correlation establishes a connection or link between log events and provides new attack concepts. Then, they analyzed the developed system's log analyzer prototype and discovered that the average accuracy of SOM #34 and AAU is 84.37% and 90.01%, respectively. The downside: More log analysis or mining must be done to obtain meaningful information and minimize the false positive and negative results. Table 2 provides a summary list of studies we reviewed for the purpose of Security.

8.3. System Maintenance

Log analysis is typically required during system maintenance because to the intricacy of network structure.

Chen *et al.* [30] studied the issue of extracting useful patterns using temporal log data. They present a new algorithm Discovering Patterns from Temporal Sequences (DTS) algorithm for extracting sequential patterns from temporally regular sequence data. Engineers can utilize the patterns that find to well know how a network-based distributed system behaves. They apply the Minimum Description Length (MDL) concept to well-known issue of pattern implosion and take another step forward in summarizing the temporal links between neighboring events in a pattern. Tests on actual log datasets showed the method's effectiveness. Extensive tests on real-world datasets show that the suggested methodologies are capable of swiftly discovering high-quality patterns.

Cheng *et al.* [31] suggested a method for detecting anomalies using log file analysis. They extract normal patterns from log data and then do anomaly detection using Principal Component Analysis (PCA). Depending on the experimental results, they concluded that the proposed technique is a great success; this enables the technique to be devised and implemented to the real log file analysis, which makes the work of the system auditor easier. With a minimum of 66% and a high of 92.3%, the average accuracy in detecting anomalies is about 80%. Table 3 provides a summary list of studies we reviewed for the purpose of Maintenance.

9. CONCLUSION

Log files are records and track of computing events across different kinds of servers and systems. ML is a reliable solution for automatically analyzing log files. Log analysis as a ML application is a fast-emerging technique for extracting information from unstructured text log data files. This study analyzed several studies from various academic databases. They each utilized a different ML method for a different objective. We have summarized the importance, methodologies, and algorithms utilized for each element we have studied. Many of the recent publications provided models intended to forecast assaults, user behavior, and system failure to improve server and system security, marketing, and failure times. The disadvantage is that methods that discriminate between normal and abnormal data require a threshold. Selecting a correct threshold is challenging and involves prior knowledge; utilizing actual datasets in model creation; many log analyses or mining must be performed to gain significant information; and minimizing false positive and negative findings. Furthermore, due to the lack of studies on, the maintenance component requires further improvements compared to the other specified features; nonetheless, interested scholars can study it further.

REFERENCES

[1] E. Shirzad and H. Saadatfar. "Job failure prediction in hadoop based on log file analysis". *International Journal of Computers and Applications*, vol. 44, no. 3, pp. 260-269, 2022.

[2] A. U. Memon, J. R. Cordy and T. Dean. "*Log File Categorization and Anomaly Analysis Using Grammar Inference*". Queen's University, Canada, 2008.

[3] M. Siwach and S. Mann. "Anomaly detection for web log data analysis: A review". *Journal of Algebraic Statistics,* vol. 13, no. 1, pp. 129-148, 2022.

[4] H. S. Malallah, S. R. Zeebaree, R. R. Zebari, M. A. Sadeeq, Z. S. Ageed, I. M. Ibrahim, H. M. Yasin and K. J. Merceedi. "A comprehensive study of kernel (issues and concepts) in different operating systems". *Asian Journal of Research in Computer Science*, vol. 8, no. 3, pp.16-31, 2021.

[5] I. Mavridis, I and H. Karatza. "Performance evaluation of cloud-based log file analysis with apache hadoop and apache spark". *Journal of Systems and Software*, vol. 125, pp. 133-151, 2017.

[6] T. Yang and V. Agrawal. "Log file anomaly detection". *CS224d Fall*, vol. 2016, pp. 1-7, 2016.

[7] S. Khan, A. Gani, A. W. A. Wahab, M. A. Bagiwa, M. Shiraz, S. U. Khan, R. Buyya and R. Y. Zomaya. "Cloud log forensics: Foundations, state of the art, and future directions". *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, pp. 1-42, 2016.

[8] V. Chitraa and A. S. Davamani. "A survey on preprocessing methods for web usage data". *International Journal of Computer Science and Information Security*, Vol. 7, no. 3, p. 1257. 2010.

[9] R. A. Bridges, T. R. Glass-Vanderlan, M. D. Iannacone, M. S. Vincent and Q. Chen. "A survey of intrusion detection systems leveraging host data". *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1-35, 2019.

[10] H. Studiawan, F. Sohel and C. Payne. "A survey on forensic investigation of operating system logs". *Digital Investigation*, vol. 29, pp. 1-20, 2019.

[11] Available from: https://www.humio.com/glossary/log-file [Last accessed on 2022 Sep 01].

[12] S. He, P. He, Z. Chen, T. Yang, Y. Su and M. R. Lyu. "A survey on automated log analysis for reliability engineering". *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1-37, 2020.

[13] M. Kumar, M. Meenu. "Analysis of visitor's behavior from web log using web log expert tool". In 2017 *International conference of Electronics, Communication and Aerospace Technology (ICECA)*. vol. 2, Institute of Electrical and Electronics Engineers, Manhattan, New York, pp. 296-301, 2017.

[14] W. Li. "*Automatic Log Analysis Using Machine Learning: Awesome Automatic Log Analysis Version 2.0*". 2013.

[15] N. Singh, A. Jain and R. S. Raw. "Comparison analysis of web usage mining using pattern recognition techniques". *International Journal of Data Mining and Knowledge Management Process*, Vol. 3, no. 4, p. 137, 2013.

[16] M. A. Latib, S. A. Ismail, O. M. Yusop, P. Magalingam and A. Azmi. "Analysing log files for web intrusion investigation using hadoop". In: *Proceedings of the 7th International Conference on Software and Information Engineering*, pp. 12-21, 2018.

[17] J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng. "A survey of machine learning for big data processing". *EURASIP Journal on Advances in Signal Processing*, Vol. 2016, no. 1, pp. 1-16, 2016.

[18] N. Jones. "Computer science: The learning machines". *Nature*, vol. 505, no. 7482, pp. 146-148, 2014.

[19] M. A. Latib, S. A. Ismail, H. M. Sarkan and R. C. Yusoff. "Analyzıng log ın bıg data envıronment: A revıew". *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 23, pp. 17777-17784, 2015.

[20] H. Xiang. "Research on clustering algorithm based on web log mining". *Journal of Physics Conf Series*, vol. 1607, no. 1, p. 012102, 2020.

[21] J. Xu, F. Xu, F. Ma, L. Zhou, S. Jiang and Z. Rao. "Mining web usage profiles from proxy logs: User identification". In: 2021 *IEEE Conference on Dependable and Secure Computing (DSC)*. Institute of Electrical and Electronics Engineers, Manhattan, New York, pp. pp. 1-6, 2021.

[22] J. Kim, M. Park, H. Kim, S. Cho and P. Kang. "Insider threat detection based on user behavior modeling and anomaly detection algorithms". *Applied Sciences*, vol. 9, no. 19, p. 4018, 2019.

[23] P. G. Prakash and A. Jaya. "Analyzing and predicting user navigation pattern from weblogs using modified classification algorithm". *Indonesian Journal of Electrical Engineering and Computer*, vol. 11, no. 1, pp.333-340, 2018.

[24] A. Abbas, M. A. Khan, S. Latif, M. Ajaz, A. A. Shah and J. Ahmad. "A new ensemble-based intrusion detection system for internet of things". *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1805-1819, 2022.

[25] V. Zeufack, D. Kim, D. Seo and A. Lee. "An unsupervised anomaly detection framework for detecting anomalies in real time through network system's log files analysis". *High Confidence Computing*, vol. 1, no. 2, pp. 100030, 2021.

[26] Y. Li, S. Yao, R. Zhang and C. Yang. "Analyzing host security using D-S evidence theory and multisource information fusion". *International Journal of Intelligent Systems*, vol. 36, no. 2, pp. 1053-1068, 2021.

[27] D. C. Le, A. N, Zincir-Heywood and M. I. Heywood. "Analyzing data granularity levels for insider threat detection using machine learning". *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, pp. 30-44, 2020.

[28] N. Shah and A. Shankarappa. "Intelligent risk management framework for BYOD". In: 2018 *IEEE 15th International Conference on e-Business Engineering (ICEBE)*. Institute of Electrical and Electronics Engineers, Manhattan, New York, pp. 289-293, 2018.

[29] S. G Tadesse and D. E Dedefa. "Layer based log analysis for enhancing security of enterprise datacenter". *International Journal of Computer Science and Information Security*, vol. 14, no. 7, pp.158, 2016.

[30] J Chen, P Wang, S Du and W Wang. "Log pattern mining for distributed system maintenance". *Complexity*, vol. 2020, no. 2, pp. 1-12, 2020.

[31] X. Cheng and R. Wang. "Communication network anomaly detection based on log file analysis". In: *International Conference on Rough Sets and Knowledge Technology*. Springer, Cham, pp. 240-248, 2014.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Rough Set-Based Feature Selection for Predicting Diabetes Using Logistic Regression with Stochastic Gradient Decent Algorithm

Kanaan M. Kaka-Khan[1], Hoger Mahmud[2], Aras Ahmed Ali[3]

[1]Department of Information Technology, University of Human Development, Iraq, [2]Department of Information Technology, the American University of Iraq, Sulaimani, [3]University College of Goizha, Sulaymaniyah

ABSTRACT

Disease prediction and decision-making plays an important role in medical diagnosis. Research has shown that cost of disease prediction and diagnosis can be reduced by applying interdisciplinary approaches. Machine learning and data mining techniques in computer science are proven to have high potentials by interdisciplinary researchers in the field of disease prediction and diagnosis. In this research, a new approach is proposed to predict diabetes in patients. The approach utilizes stochastic gradient descent which is a machine learning technique to perform logistic regression on a dataset. The dataset is populated with eight original variables (features) collected from patients before being diagnosed with diabetes. The features are used as input values in the proposed approach to predict diabetes in the patients. To examine the effect of having the right variable in the process of making predictions, five variables are selected from the dataset based on rough set theory (RST). The proposed approach is applied again but this time on the selected features to predict diabetes in the patients. The results obtained from both applications have been documented and compared as part of the approach evaluations. The results show that the proposed approach improves the accuracy of predicting diabetes when RST is used to select variables for making the prediction. This paper contributes toward the ongoing efforts to find innovative ways to improve the prediction of diabetes in patients.

Index Terms: Logistic Regression, Stochastic Gradient Descent, Rough Set Theory, K-fold Cross-validation, Diabetes Prediction

1. INTRODUCTION

Changes in human lifestyle and the deterioration of the environment have left a negative impact on human health. For that reason, human health has always been the subject of research with the aim to improve it. Diabetes is a group of metabolic diseases which result in high blood sugar levels for a prolonged period. As stated by International Diabetes Federation, 537 million adults (20–79 years) are living with diabetes which is 1 in 10 of adult population. This number is predicted to rise to 643 million by 2030 and 783 million by 2045 [1]. Diabetes has been the subject of research for some times by multidisciplinary scientists with the aim to find and improve methods that lead to effective prevention, diagnosis, and treatment of the disease. For instance, in a similar approach, in 2013, Anouncia *et al.* proposed a diagnosis system for diabetes. The system is implemented to diagnose the type of diabetes based on symptoms provided by patients. They have used rough set-based knowledge representation in developing their system and the results showed improvements in terms of accuracy of diabetes type diagnosis and the time it takes for the diagnosis [2]. Despite all the efforts invested into researching diagnostic techniques for diabetes, research

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.pp85-93

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Kanaan M. Kaka-Khan, Department of Information Technology, University of Human Development, Iraq. E-mail: kanaan.mikael@uhd.edu.iq

Received: 21-08-2022 Accepted: 02-10-2022 Published: 18-10-2022

shows that there is still room for improvement, especially in areas related to the level of accurately in predicting the disease in a patient. Rough set theory (RST) has been used by researchers to predict a wide array of topics such as time series prediction [3], crop prediction [4], currency crisis prediction [5], and stock market trends prediction [6]. In this research, we use RST to select variables in a dataset with the aim to improve the level of accuracy in predicting diabetes in a patient. Stochastic gradient descent algorithm is used to process the variables selected to make diabetes prediction based on computed logistic regression values from the dataset. The dataset used for all experiments in this study is made available by the Pima Indian Diabetes [7]. This paper contributes toward the ongoing efforts to find innovative ways to improve the prediction of diabetes in patients by proposing a new approach to predict diabetes in patients using machine learning techniques. The results presented in Sections 5.1 and 5.2 show that the approach improves accuracy in making diabetes predictions compared to other available approaches.

The rest of this paper is organized as follows: Section 2 provides the theoretical background needed to understand the selected techniques and Section 3 provides a survey of related literatures. Section 4 provides the description of the methodology used in this study. Experimental results and discussion are provided in Section 5. Finally, conclusions are drawn in Section 6.

2. BACKGROUND

This section provides a basic background on the theories used in the study.

2.1. RST

Rough set [8] is proposed by Pawlak to deal with uncertainty and incompleteness. It offers mathematical tools to discover patterns hidden in datasets and identifies partial or total dependencies in a dataset based on indiscernibility relation. The technique calculates a selection of features to determine the relevant feature. The general procedures in rough set are as follows:

The Lower Approximation of set D is the set of objects in a table of information which certainly belongs to the class X:

$$AX = \{xi \in U \mid [xi]_{ind(A)} \subset X\}, \text{X} \in \text{Att} \quad (1)$$

The Upper Approximation of a set X includes all objects in a table of information which possibly belongs to the class X:

$$AX = \{xi \in U \mid [xi]_{(A)} \cap A \neq \phi\} \quad (2)$$

Boundary Region is the difference between upper approximation set and lower approximation set that is referred to as Bnd (X)

$$\beta = AX - \underline{AX} \quad (3)$$

Positive Region is the set of all objects that belong to lower approximation, which means, the union of the lower approximation consist of the union of all the lower approximation sets:

$$\rho = \cup \underline{A} \text{ (Union of all lower sets)} \quad (4)$$

Indiscernibility of positive reign for any G $\subseteq$ Att is the associated equivalence relation:

$$\text{IND (G)} = \{(x, y) \in p\times : \forall a \in \text{G}, \alpha (x) = (y)\} \quad (5)$$

Reducts are the minimum range representation of the original data without loss of information:

$$\text{reducts } \delta = \min IND \quad (6)$$

2.2. Stochastic Gradient Descent

According to [9], stochastic gradient descent is a function's minimizing process, following the slope or gradient of that function. In general, in machine learning, stochastic gradient descent can be considered as a technique to evaluate and update the weights every iteration, which minimizes the error in training data models. While training, this optimization technique tries to show each and every training sample to the model one by one. For each training sample, the model produces an output (prediction), calculates the error, and updates to minimize the error for the next output, and this process is repeated for a fixed number of epochs or iterations. Equation-7 describes the way of finding and updating the set of weights (coefficients) in a model from the training data.

$$\text{B} = \text{b-learning rate} \times \text{error} \times x \quad (7)$$

Here, b is the coefficient (weight) being estimated, learning rate is a learning value that can be configured between (0.01 and 10), error is the model's predicted error, and x is the input value. The accuracy of the prediction can be calculated simply by dividing the number of corrected predictions by the actual values produced in formula 3.

$$Accuracy = \frac{\sum Correct\ predictions}{\sum Actual\ values} \quad (8)$$

2.3. Logistic Regression

Logistic regression [10] is a two-class problems linear classification algorithm. Equation 9 represents the logistic regression algorithm. In this algorithm, to make a prediction (*y*), using coefficient (weight) values, the input values (*X*) are combined in a linear form. Logistic regression produces an output of binary value (0 or 1).

$$yhat = \frac{1.0}{1.0 + e^{-(b0+b1\times x1)}} \quad (9)$$

The foundation of logistic regression algorithm is Euler's number, the estimated output is represented as *yhat*, the algorithm's bias is b_0, and the coefficient (weight) for the single input value (x_1) is represented as b_1. The logistic regression produces a real value as an output (yhat) which is between 0 and1. To be mapped to an estimated class value, the output needs to be converted (rounded) to an integer value. Each column (attribute) of the dataset has an associate value (b) that should be estimated from the training data and it is the actual model's representation that can be saved for further use.

3. RELATED WORK

Prediction is a widely used approach in many fields of science including healthcare to foresee possible outcomes of a cause. Disease prediction is certainly an area, where researchers have been working by applying a number of different theories including machine learning theories with the aim to find methods to make the most accurate prediction possible. RST is one of the theories used to classify and predict diseases. For instances, the authors of [11] have used the theory to classify medical diagnosis, the authors of [12] and [13] have modified and used the theory to improve disease prediction. Type 1 and 2 diabetes were the focus of the authors of [14], in which they developed a hybrid reasoning model to address prediction accuracy issues. Based on their results, they claim that their approach raises diabetes prediction accuracy to 95% compared to other existing approaches. In 2017, RST was used by the authors of [15] to develop a model for patient clustering in a dataset. The authors considered average values calculated from diabetes indicators in a dataset to cluster the patients in it. In the same year, deep learning was utilized by the authors of [16] to establish an intelligent diabetes prediction model, in which patients' risk factors collected in a dataset were considered to make the prediction.

In 2018, Fuzzy RST is applied first to select specific features in a dataset, later in the process, to improve prediction performance, save processing time, and better diagnosis accuracy that the Optimized Generic Algorithm (OGA) is applied. The results obtained from the study shows that the approach has achieved the objectives of the study [17]. In 2020, Vamsidhar Talasila and Kotakonda Madhubabu proposed the use of RST technique to select the most relevant features to be inputted to the Recurrent Neural Network (RNN) technique for disease prediction. They claimed that the RST-RNN method achieved accuracy of 98.57% [18]. In the same year, Gao and Cheng proposed an improved neighborhood rough set attribute reduction algorithm (INRS) to increase the dependence of conditional attributes based on considering the importance of individual features for diabetes prediction [14]. In 2021, Gadekallu and Gao proposed a model using an approach based on rough sets to reduce the attributes needed in heart disease and diabetes prediction [19]. The main limitation of these studies is the fact that none has considered the quantity and quality of viables used to make diagnostic predictions.

The approach used in this study is similar to the ones used in the surveyed literatures but differs in objectives. We use RST to select the best features in a dataset and use stochastic gradient decent algorithm to compute the logistic regression values from the selected features in the dataset with the aim to improve the prediction accuracy of diabetes in a patient.

4. METHODOLOGY

This section provides insights on the methodology used to achieve the objectives of the study. The methodology is comprised six major steps:

4.1. Step 1

A dataset is selected, examined for suitability and reliability based on a number of characteristics, and uploaded to be analyzed. The dataset selected and uploaded for the purpose of this research is provided by Pima Indians Diabetes [7]. The selected dataset involves predicting diabetes within 5 years in Pima Indians given medical details. The dataset is a 2-class classification problem and consists of 76 samples with 8 input and 1 output variable. The variable names are as follows: Number of Times Pregnant, Plasma Glucose concentration a 2 h in an oral glucose tolerance test, Diastolic Blood Pressure (mm Hg), Triceps Skinfold Thickness (mm), 2-h Serum Insulin (mu U/ml), Body Mass Index (weight in kg/[height in m]^2), Diabetes Pedigree Function, Age, and Class Variable (0 or 1). Before implementing the model, it is highly preferred to do preprocessing due to some

deficiencies. Usually, the dataset contains features highly varying in magnitudes, units, and range which may results in inaccurate output [20]. In this work due to use of stochastic gradient descent algorithm, the dataset has been normalized using min-max scaling to bring all values to between 0 and 1. Table 1 shows a sample of the selected dataset.

4.2. Step 2

The selected diabetes dataset is preprocessed and normalized. To increase the efficiency and accuracy of the model, the dataset needs to be pre-processed before applying the proposed model since the data may contain null values, incorrect, and redundant information. In general, data processing involves two major steps: data cleaning and data normalization. Data cleaning means removing incorrect information or filling out missing values to increases the validity and quality of a dataset though applying a number of different methods [21]. In this study, in case of any tuple containing missing values, the missed attribute value assumed to be 0 (this is achieved using the fill_mising_values () function from the python script developed for the implementation phase of this study). Redundant or unnecessary columns are deleted to have a high quality dataset (this is achieved using the remove_duplicate_columns () function from the python script). To let all features have equal weight and contribution to the model, the range of each feature needs to be scaled, for this purpose, the dataset is normalized to a range of [0,1] by the following processes: *String columns converting:* the string columns are converted to float through *str* column using the float() function. *Min max finding:* min and max values of each column of the dataset are found through using the dataset minmax() function. Finally, the dataset is normalized by the min-max normalization method using the following equation adapted form [22].

$$X' = \frac{x - min(x)}{max(x) - min(x)} \quad (10)$$

4.3. Step 3

In this step, RST is applied to select the features which might produce a better prediction. There are *nine* variables in total in the dataset, as shown in Table 1. The class variable is considered as a dependent variable and the other eight variables are assumed as predictors or independent variables. Table 2 presents the regression calculation summary for diabetes classification of the dataset. The result of the calculation clearly shows that the accuracy of diabetes prediction is *30.32%* if all variables in the dataset are considered in the calculation. The low accuracy result is an indication that there might be one or more variables which are not fit to be used for prediction. The regression calculation also shows that the un-standardized regression coefficient (*b*) is 0.06 for pregnancies, which indicates that if all other predictors are controlled then an increment of one unit in pregnancies increases the accuracy by 0.06. The same statement can be made for the other variables. To flitter the features that might produce a better diabetes prediction, the dataset is grouped together into nine elementary sets based on indiscernibility relation level between the data elements. Table 3 shows the details of the groups. To further process the groups, the discernibility matrix has been developed for the elementary sets and the result is shown in Table 4. From the discernibility matrix, a discernibility function has been developed, as shown in equation 11.

$$f(A) = f(A1) \times f(A2) \times \ldots \times f(An) \quad (11)$$

As the result of discernibility function of all elementary sets for the entire dataset, we found that:

$f(A) = a1 \vee a2 \vee a5 \vee a6 \vee a8$ where a1 is Pregnancies; a2 is Plasma glucose; a5 is Insulin; a6 is DPF; and a8 is age attribute. Table 5 shows the reduct matrix for the elementary sets. From the reduct matrix, all reducts and core attributes have been found:

TABLE 1: The first ten records of the diabetes dataset used in this study

Pregnancies	Plasma glucose	Blood pressure	Skinfold thickness	Insulin	BMI	DPF	Age	Class variable
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

BMI: Body mass index

f(R1) = a1∨a2∨a6; f(R2) = a1∨a2∨a5∨a8; f(R3) = a2∨a5∨a8; f(R4) = a1∨a2∨a8; f(R5) = a2∨a6∨a8; f(R6) = a1∨a2∨a6∨a8; f(R7) = a2∨a5∨a6; f(R8) = a1∨a2∨a5; f(R9) = ∨a2∨a5∨a6∨a8. Finally, Table 6 shows the features that are selected to be used for making diabetes prediction.

TABLE 2: Linear regression statistics of diabetes dataset

Multiple R			0.550684207		
R Square			0.303253096		
Adjusted R Square			0.295909255		
Standard Error			0.400210451		
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Unstandardized regression coefficient (b)**
Intercept	0.853894266	0.085484958	-9.98882	0.00	0.066
Pregnancies	0.020591872	0.00512998	4.014026	0.00	1.863
Plasma glucose	0.005920273	0.000515123	11.49294	0.00	0.022
blood pressure	0.002331879	0.000811639	2.87305	0.00	0.081
Skinfold thickness	0.00015452	0.001112215	0.13893	0.89	0.247
Insulin	0.000180535	0.000149819	-1.20502	0.23	0.004
MI	0.013244031	0.00208776	6.343656	0.00	0.000
DPF	0.147237439	0.045053885	3.26803	0.00	0.686
Age	0.002621394	0.00154864	1.692707	0.09	0.001

TABLE 3: Elementary sets

Samples	**Pregnancies**	**Plasma glucose**	**Blood pressure**	**Skinfold thickness**	**Insulin**	**BMI**	**DPF**	**Age**
Group 1	0–1	0–22	0–13	0–10	0–94	0–6	0–0.25	21–26
Group 2	2–3	23–46	14–28	11–22	95–190	7–14	0.26–0.51	27–33
Group 3	4–5	47–70	29–43	23–34	191–286	15–22	0.52–0.77	34–41
Group 4	6–7	71–94	44–58	35–46	287–382	23–30	0.78–1.03	42–49
Group 5	8–9	95–118	59–73	47–58	383–478	31–38	1.04–1.29	50–57
Group 6	10–11	119–142	74–88	59–70	479–574	39–46	1.3–1.55	58–63
Group 7	12–13	143–166	89–103	71–82	575–670	47–54	1.56–1.81	64–69
Group 8	14–15	167–190	104–118	83–94	671–766	55–62	1.82–2.03	70–75
Group 9	16–17	191–199	119–122	95–99	767–846	63–67	2.04–2.42	76–81

TABLE 4: Discernibility matrix

Samples	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**	**Group 7**	**Group 8**	**Group 9**
Group 1	-								
Group 2	a1a2a4a7a8	-							
Group 3	a2a3a4a8	a1a3a4a8	-						
Group 4	a1a2a4a6a7	a2a3a4a7a8	a1a2a7a8	-					
Group 5	a2a3a5a7a8	a1a3a4a8	a1a2a4a6a7	a2a3a5a7a8	-				
Group 6	a1a3a5a6a8	a3a4a6a8	a1a3a5a7a8	a2a4a5a7a8	a2a3a5a7a8	-			
Group 7	a1a2a4a6a8	a2a4a5a7	a1a2a4a6	a2a3a5a7a8	a2a3a5a8	a5a6a7	-		
Group 8	a1a2a4a6a7	a1a3a4a8	a1a2a7a8	a2a3a5a7a8	a2a4a5a7	a3a4a5	a2a4a5	-	
Group 9	a2a4a5a7	a1a2a4a7	a3a5a8	a2a5a7a8	a3a4a6	a2a3a8	a2a4a5	a3a4a5	-

TABLE 5: Reducts matrix

Samples	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**	**Group 7**	**Group 8**	**Group 9**
Group 1	-	a1a2a4a7a8	a2a3a4a8	a1a2a4a6a7	a2a3a5a7a8	a1a3a5a6a8	a1a2a4a6a8	a1a2a4a6a7	a2a4a5a7
Group 2	a1a2a4a7a8	-	a1a3a4a8	a2a3a4a7a8	a1a3a4a8	a3a4a6a8	a2a4a5a7	a1a3a4a8	a1a2a4a7
Group 3	a2a3a4a8	a1a3a4a8	-	a1a2a7a8	a1a2a4a6a7	a1a3a5a7a8	a1a2a4a6	a1a2a7a8	a3a5a8
Group 4	a1a2a4a6a7	a2a3a4a7a8	a1a2a7a8	-	a2a3a5a7a8	a2a4a5a7a8	a2a3a5a7a8	a2a3a5a7a8	a2a5a7a8
Group 5	a2a3a5a7a8	a1a3a4a8	a1a2a4a6a7	a2a3a5a7a8	-	a2a3a5a7a8	a2a3a5a8	a2a4a5a7	a3a4a6
Group 6	a1a3a5a6a8	a3a4a6a8	a1a3a5a7a8	a2a4a5a7a8	a2a3a5a7a8	-	a5a6a7	a3a4a5	a2a3a8
Group 7	a1a2a4a6a8	a2a4a5a7	a1a2a4a6	a2a3a5a7a8	a2a3a5a8	a5a6a7	-	a2a4a5	a2a4a5
Group 8	a1a2a4a6a7	a1a3a4a8	a1a2a7a8	a2a3a5a7a8	a2a4a5a7	a3a4a5	a2a4a5	-	a3a4a5
Group 9	a2a4a5a7	a1a2a4a7	a3a5a8	a2a5a7a8	a3a4a6	a2a3a8	a2a4a5	a3a4a5	-

TABLE 6: Indiscernibility table

Samples	Pregnancies	Plasma glucose	Insulin	DPF	Age
Group 1	0–1	0–22	*	0–0.25	*
Group 2	2–3	23–46	95–190	*	27–33
Group 3	*	47–70	191–286	0.52–0.77	34–41
Group 4	*	71–94	287–382	*	42–49
Group 5	8–9	95–118	*	*	50–57
Group 6	*	119–142	*	1.3–1.55	58–63
Group 7	12–13	143–166	*	1.56–1.81	64–69
Group 8	14–15	167–190	671–766	*	*
Group 9	*	191–199	767–846	2.04–2.42	76–81

Table 3 shows the indiscernibility level of the relation between the patients.

Table 6 represents the last step of RST process, in which the data are simplified, and the indiscernibility relations are stated. The * symbol means that a certain variable has no impact in a certain case, for example, if the patient's pregnancy is (0–1) and plasma glucose is (0–22) and DPF is (0-0.25), then the patient has diabetes regardless of the value of other attributes, and so on.

4.4. Step 4

In this step, the logistic regression algorithm with stochastic gradient descent technique is applied on the selected features in the previous step. The major steps of the application are as follows:

4.4.1. Dataset loading

The dataset is loaded into the model through load_dataset() function.

4.4.2. Dataset preprocessing

The dataset is preprocessed through str column to float(), dataset minmax(), and normalize dataset() functions accordingly.

4.4.3. Dataset splitting into k folds

The dataset is split into k-folds and trainset. Test set creation for training the model is achieved through cross validation split() function.

4.4.4. Coefficients estimating

Coefficients or weights are the values that determine the model accuracy and can be estimated for training data using stochastic gradient descent. The algorithm uses two parameters to estimate the weights (coefficient), the first one is learning rate to specify the amount of each weight, and it is corrected continuously, while it is updated. The second one is Epochs which is the loop through the training process while updating the coefficient. The Coefficients Estimating *is achieved through* coefficients sgd() *function.*

4.4.5. Coefficients updating

For each instance in the training data, each coefficient is updated throughout all epochs. The error that the model makes is the criteria for updating the coefficients. The simple equation can be used to calculate the error (equation-12).

Error = (Expected output value) – (Prediction made with the candidate coefficients) (12)

4.5. Step 5

Predictions are generated; equation 7 describes the prediction process which is the most important part of the model. Prediction process will be needed twice: first in stochastic gradient descent to evaluate candidate coefficient values and second in the model when it is finalized to produce outputs (predictions) on test data. The prediction process is achieved through predict() function. Fig. 1 shows the execution flow of the proposed approach.

4.6. Step 6

Finally, the results obtained are compared. Fig. 1 shows the proposed diabetes prediction method.

4.7. Model Performance Evaluation

In this research, k-fold cross-validation technique has been used to evaluate the learned model's performance on unseen data. Cross-validation is a resampling procedure used to validate machine learning models on a limited data sample. Using k-fold, cross-validation means that k models will be construct, evaluated, and through using mean model error, the model's performance is estimated. After rounding the predicted value of each row which is a float number between 0 and 1, it will be compared to its actual value. If they are equal, the prediction is considered as a correct result. Simple error equation (equation 13) will be used to evaluate each model.

$$Accuracy = \frac{No.\ of\ correct\ results}{Total\ no.\ of\ samples} * 100 \quad (13)$$

The general procedure is as follows: (1) Shuffle the dataset randomly. (2) Split the dataset into k groups, (3) take a group as a test set and the remaining as a training set, the same procedure will be repeated for each and every group; (4) as usual, the model will be Fitting on the training set and evaluating on the test set, and (5) retain the result (evaluation score) the model can be discarded [17], [23]. For this work,

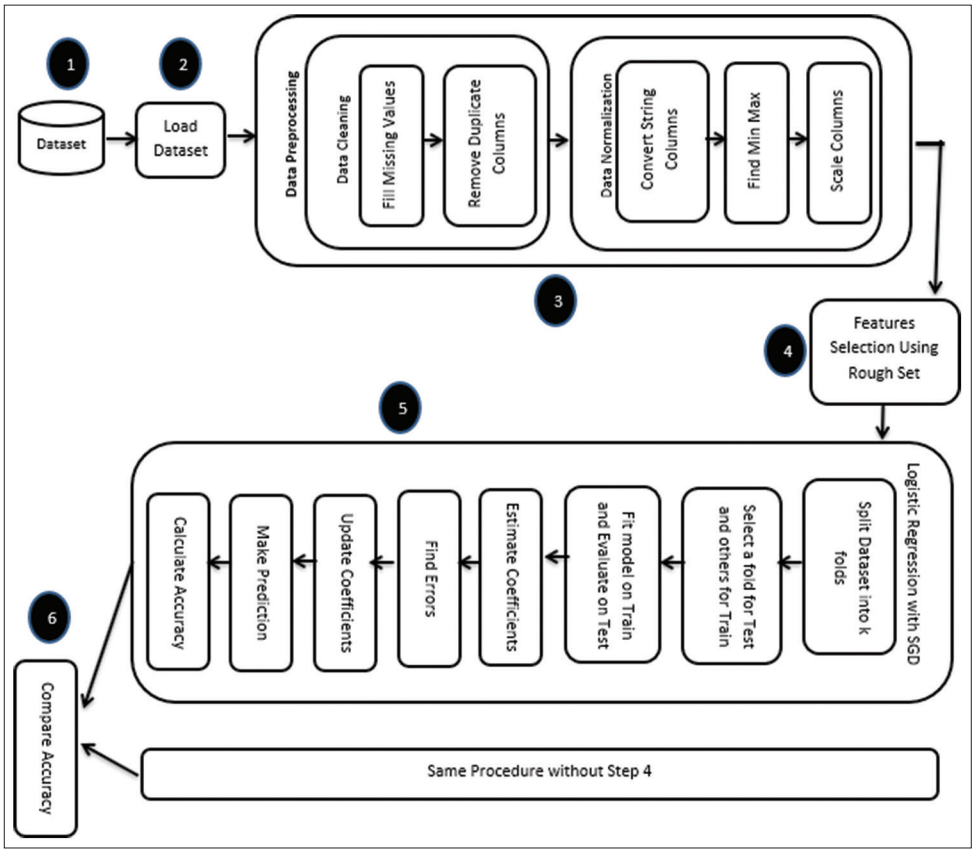


Fig. 1. Proposed diabetes prediction method.

a learning rate, training epochs, and k value are (0.1, 100, 5) subsequently.

After implementing the model twice; first on the dataset with all features, and second with features selected by applying RST, the results can be discussed as follows:

4.8. Making Prediction on Dataset with all Features

The aim of using logistic regression is predicting the dependent variable (output variable) based on equation 7, and the aim of using stochastic gradient descent technique is minimizing the error of predicted coefficient values while training the model on the dataset. For model training, k-fold cross-validation technique is used to split out the dataset to 5 folds (groups), a fold is used as a test set and the others as train sets, for example:

- Mode l: Fold1 for test and fold2, fold3, fold4, and fold5 for train
- Mode 2: Fold2 for test and fold1, fold3, fold5, and fold5 for train
- Mode 3: Fold3 for test and fold1, fold2, fold4, and fold5 for train
- Mode 4: Fold4 for test and fold1, fold2, fold3, and fold5 for train
- Mode 5: Fold5 for test and fold1, fold2, fold3, and fold4 for train.

TABLE 7: Accuracy score of each model used

Model No.	Accuracy
Model 1	73.857
Model 2	78.431
Model 3	81.699
Model 4	75.816
Model 5	75.816
Score	77.124%

For each model, after training for 100 epochs (iterations) and minimizing the errors to a desired results and calculate the accuracy using equation 11, the score can be calculated using equation 14.

$$Score = \frac{Sum\ of\ all\ model\ acuracy\ results}{Total\ no.of\ models} \quad (14)$$

The total number of models used is five. Table 7 summarizes the models result and the overall score. The overall score is 77.12% for the model on the dataset with all features.

4.9. Making Prediction on Dataset with RST-Based Selected Feature

The same process applied on the dataset with selected features based on RST, the result is presented in Table 8.

TABLE 8: Accuracy and score for all five models for selected features

Model No.	Accuracy
Model 1	77.342
Model 2	81.013
Model 3	83.874
Model 4	78.394
Model 5	79.628
Score	80.215%

Table 9 shows the comparison between the results obtained from both implementations; implementing the model on the dataset with all features and the RST-based selected features. The results show that RST-based selected features for machine learning compared to the data set with all features give more accurate predictions.

TABLE 9: Accuracy and score for all five models using all features, RST-based selected features

Model No.	All features (Accuracy)	RST-based selected features (Accuracy)
Model 1	73.856	77.342
Model 2	78.431	81.013
Model 3	81.699	83.874
Model 4	75.816	78.394
Model 5	75.816	79.628
Score	77.124%	80.215%

RST: Rough set theory

The baseline score for the selected dataset is 65% our experiment results which indicated that the proposed approach increased the prediction accuracy for diabetes dataset with all features from 65% to 77% and 80% for RST-based features dataset, as shown in Table 10.

TABLE 10: Accuracy summery of baseline and proposed algorithm for diabetes

Model name	Prediction accuracy (%)
Baseline score	65
Logistic regression with SGD algorithm	77.124
RST-based logistic regression with SGD algorithm	80.215

Finally, it can be summarized that implementing the logistic regression algorithm with stochastic gradient descent technique is one of the suitable choices for diabetes predictions on the basis of the results. At the same time, rather than using all features, more precise predictions can be made by feature selection based on rough set for neural network. Table 11 summarizes a comparison between our works with some of the most recently published works.

TABLE 11: Dataset classification comparison

Works	Data size	Methods	Accuracy (%)
[24]	768 samples with 9 attributes	Logistic Regression	77
[25]	768 samples with 9 attributes	Modified PSO Naïve Bayes	78.6
[26]	768 samples with 9 attributes	Modified Weighted knn (SDKNN)	83.76
[27]	768 samples with 9 attributes	random forest classifier	79.57
Our proposed method	768 samples with 9 attributes	Logistic regression with SGD algorithm	77
	768 samples with 6 attributes	RST-based logistic regression with SGD algorithm	80.215

RST: Rough set theory

5. CONCLUSION AND FUTURE WORK

In the health-care sector predicting, the presence or non-presence of diseases is important to help people know their health status so that they take the necessary steps to control the disease.

This paper explores the use of stochastic gradient descent algorithm to apply logistic regression on datasets to make predictions on the presence of diabetes. The Pima Indian Diabetes dataset is used to produce results using the proposed technique. The experiments results show that diabetes can be predicted more accurately using logistic regression with stochastic gradient descent algorithm when RST is used to select the important features on a normalized dataset. This is paper makes a real contribution in the use of interdisciplinary techniques to improve prediction mechanisms in health-care sector in general diabetes prediction in specific. The main purpose of this work is showing the significance of using RST with machine learning algorithms, hence in the future; the same theory can be applied with other algorithms to have a better result.

REFERENCES

[1] "Diabetesatlas". Available from: https://www.diabetesatlas.org [Last accessed on 2022 Aug 08].

[2] M. Anouncia, C. Maddona, P. Jeevitha and R. Nandhini. "Design of a diabetic diagnosis system using rough sets". *Cybernetics and Information Technologies*, vol. 13, no. 3, pp. 124-169, 2013.

[3] F. E. Gmati, S. Chakhar, W. L. Chaari and H. Chen. "A rough set approach to events prediction in multiple time series". In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, vol. 10868, pp. 796-807, 2018.

[4] H. Patel and D. Patel. "Crop prediction framework using rough set theory". *International Journal of Engineering and Technology*, vol. 9, pp. 2505-2513, 2017.

[5] S. K. Manga. "Currency crisis prediction by using rough set theory". *International Journal of Computer Applications*, vol. 32, p. 48-52, 2011.

[6] B. B. Nair, V. Mohandas and N. Sakthivel. "A decision tree-rough set hybrid system for stock market trend prediction". *International Journal of Computer Applications*, vol. 6, no. 9, pp. 1-6, 2010.

[7] "Pima-Indians-Diabetes-Dataset". Available from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Last accessed on 2022 May 04].

[8] Z. Pawlak. "Rough set theory and its applications to data analysis". *Cybernetics and Systems*, vol. 29, no. 7, pp. 661-688, 1998.

[9] P. Achlioptas. "*Stochastic Gradient Descent in Theory and Practice*". Stanford University, Stanford, CA, 2019.

[10] J. Brownlee. *Machine Learning Algorithms from Scratch with Python*. Machine Learning Mastery, 151 Calle de San Francisco, US, 2016.

[11] H. H. Inbarani and S. U. Kumar. "A novel neighborhood rough set based classification approach for medical diagnosis". *Procedia Computer Science*, vol. 47, pp. 351-359, 2015.

[12] E. S. Al-Shamery and A. A. R. Al-Obaidi. "Disease prediction improvement based on modified rough set and most common decision tree". *Journal of Engineering and Applied Sciences*, vol. 13, no. Special issue 5. pp. 4609-4615, 2018.

[13] R. Ghorbani and R. Ghousi. "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction". *International Journal of Data and Network Science*, vol. 3, no. 2, pp. 47-70, 2019.

[14] R. Ali, J. Hussain, M. H. Siddiqi, M. Hussain and S. Lee. "H2RM: A hybrid rough set reasoning model for prediction and management of diabetes mellitus". *Sensors*, vol. 15, no. 7, pp. 15921-15951, 2015.

[15] S. Sawa, R. D. Caytiles and N. C. S. Iyengar. "A Rough Set Theory Approach to Diabetes". In: *Conference: Next Generation Computer and Information Technology*, 2017.

[16] S. Ramesh, H. Balaji, N. Iyengar and R. D. Caytiles. "Optimal predictive analytics of pima diabetics using deep learning". *International Journal of Database Theory and Application*, vol. 10, no. 9, pp. 47-62, 2017.

[17] K. Thangadurai and N. Nandhini. "Integration of rough set theory and genetic algorithm for optimal feature subset selection on diabetic diagnosis". *ICTACT Journal on Soft Computing*, vol. 8, no. 2, 2018.

[18] V. Talasila, K. Madhubabu, K. Madhubabu, M. Mahadasyam, N. Atchala and L. Kande. "The prediction of diseases using rough set theory with recurrent neural network in big data analytics". *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 10-18, 2020.

[19] T. R. Gadekallu and X. Z. Gao. "An efficient attribute reduction and fuzzy logic classifier for heart disease and diabetes prediction". *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 14, no. 1, pp. 158-165, 2021.

[20] "Medium". Available from: https://www.medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e [Last accessed on 2022 Jun 05].

[21] E. Rahm and H. H. Do. "Data cleaning: Problems and current approaches". *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3-13, 2000.

[22] D. Borkin, A. Némethová, G. Michal'conok and K. Maiorov. "Impact of data normalization on classification model accuracy". *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, vol. 27, no. 45, pp. 79-84, 2019.

[23] "Machine Learning Mastery". Available from: https://www.machinelearningmastery.com/k-fold-cross-validation [Last accessed on 2022 Aug 06].

[24] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta and S. K. Tayebati. "Comparative machine-learning approach: A follow-up study on Type 2 diabetes predictions by cross-validation methods". *Machines*, vol. 7, no. 4, pp. 74, 2019.

[25] D. K. Choubey, P. Kumar, S. Tripathi and S. Kumar. Performance evaluation of classification methods with PCA and PSO for diabetes. *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, p. 5, 2020.

[26] R. Patra and B. Khuntia. "Analysis and prediction of Pima Indian diabetes dataset using SDKNN classifier technique". *IOP Conference Series: Materials Science and Engineering*, vol. 1070, no. 1, p. 012059, 2021.

[27] V. Chang, J. Bailey, Q. A. Xu and Z. Sun. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms". *Neural Computing and Applications*. vol. 34, no. 10, pp. 1-7, 2022.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Secure Medical Image Transmission System Based on 2D Logistic Map and Diffie–Hellman Key Exchange Mechanisms

Shakhawan H. Wady[1,2], Raghad Z. Yousif[3]

[1]Department of Business Administration, Business College, Charmo University, Chamchamal, Sulaimani, KRG, Iraq, [2]Department of Information Technology, University College of Goizha, Sulaimani, KRG, Iraq, [3]Department of Physics, College of Science, Salahaddin University, Erbil, KRG, Iraq

ABSTRACT

With the tremendous growth of searchable visual media, the content-based medical image retrieval and computer-aided diagnosis systems have become popular in recent years to improve knowledge and provide facilities for radiologists. Medical images transferred throughout public networks demand a mechanism that guarantees image privacy, ownership and source of origin reliability, and image integrity verification. For this reason, secure image retrieval and diagnosis scheme have been given considerable interest due to users' security concerns. This work proposed a secure framework based on a two-dimensional (2D) chaotic map with Diffie–Hellman key exchange protocols to ensure patient information privacy and security. Consequently, from a security and protection perspective, the objective is to provide a privacy procedure for medical image retrieval systems through image encryption technique combined with a secure key exchange procedure to minimize the possibility of secret key interception by an unauthorized person. Simulation results and security analysis show that the suggested technique could protect images with minimal time complexity and a high level of security while also resisting numerous attacks.

Index Terms: Medical Images, Chaotic Map, Diffie–Hellman Key Exchange, Peak Signal-to-Noise Ratio

1. INTRODUCTION

Due to the extensive transmission of medical image data through several communication networks, security concerns have become more and more prominent. The protection of personal information is difficult to guarantee by relying exclusively on access control [1], [2]. Therefore, developing good image security systems have become a focal research topic and attract extensive public and government concerns. An encryption procedure is a mechanism, in which the user transforms the plain text or image into an unintelligible form called a cipher or a cipher image. While decryption is a reverse process of encryption, in which the cipher image is converted back into plain text image (Original image) with the help of a key sequence. Symmetric and asymmetric key encryption techniques are the two important categories of encryption procedures through which the medical images are encrypted and decrypted [3], [4]. The technique of chaos-based image encryption is considered a good candidate for cryptography among numerous approaches for content-based image retrieval (CBIR) and computer-aided diagnosis (CAD) systems on large image collections. In consequence, using the chaos scheme for image encryption and decryption is superior than existing conventional algorithms. Since Fridrich [5], [6] previously recommended a chaos-based encryption mechanism in 1997, chaos-based image encryption algorithms, including the Henon map, Baker map, logistic map, and Arnold cat map, have been applied

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp94-104

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: shakhawan.hares@charmouniversity.org

Received: 26-04-2022 Accepted: 28-09-2022 Published: 27-10-2022

in numerous literature-based data encryption [7]–[9]. Most chaos-based cryptosystems were starting to use it as a core structure. The Fridrich encryption technique is made up of two layers: A confusion layer that uses the 2-D Baker chaotic map and a diffusion layer.

This work provides a novel security system operation platform to guarantee the safety of the medical images based on chaotic cryptography along with Diffie–Hellman key exchange scheme. The key contribution of this study is to utilize chaotic cryptography to terminate the intelligibility of all the retrieved medical images and applying Diffie–Hellman key exchange protocol to eliminate the need for sending the encryption key into private secure channel. The rest of the paper is organized as follows. Section 2 puts forward a literature review. Section 3 presents a complete architecture overview of the proposed system operation scenario, including sections such as system architecture, encryption and uploading, and downloading and decryption stages. Section 4 discusses the results and analysis. Finally, Section 5 provides the conclusion of the work.

2. LITERATURE REVIEW

In this section, necessary background has been presented on the uses of two-dimensional logistic map and Diffie–Hellman key exchange mechanism in image encryption schemes and literature survey medical image encryption schemes proposed in the literature.

2.1. Two-Dimensional Logistic Map

The two-dimensional logistic map is a discrete dynamic system with a chaotic behavior including less periodic windows in bifurcation diagrams and a larger range of parameters, which are more suitable for cryptography [10], [11]. It has a perfect chaotic property as a traditional logistic algorithm, and it has more complex behavior than one dimensional chaotic behavior. Mathematically, the two-dimensional logistic map is an example for chaotic map, and it can be discretely defined using Equation (1) as follows [12]:

$$2D\ Logistic\ map:$$

$$Z(x,y)=\begin{Bmatrix} x_{i+1}=r(3y_i+1)x_i\ (1-x_i) \\ y_{i+1}=r(3x_{i+1}+1)y_i\ (1-y_i) \end{Bmatrix} \quad (1)$$

Where:

- $Z(x, y)$ is the 2D logistic map
- r is the control parameter (growth rate)
- (x_i, y_i) is the pairwise point at the i^{th} iteration.

The 2D logistic map defined in Equation (1) is more complex than the 1D logistic map, that is, the conventional logistic map is defined in Equation (2), where r is the chaotic behavior control parameter [13].

$$\text{1D Logistic map: } x_{i+1} = rx_i(1-x_i) \quad (2)$$

Fig. 1 shows the 1D logistic map schema, where the horizontal axis is the growth rate which is indicated by the parameter (r) and vertical axis is the population which is indicated by the parameter (x) and the trajectories of every one-dimensional logistics map are shown by the (x) with a fixed (x) as points on Figure.

In this work, 2D logistic map algorithm was performed to encrypt and decrypt the medical images with high security, which facilitates the process of protecting the private information of medical images.

2.2. Diffie–Hellman Key Exchange Mechanism

The Diffie–Hellman key agreement protocol, also known as an exponential key agreement, is a symmetric key exchange protocol which is widely implemented for encryption process, as shown in Fig. 2. The main objective of the Diffie–Hellman procedure is to make it feasible for two or more parties to build and exchange an identical and hidden session key by exchanging information over a public communications channel [14]. The DH key agreement protocol allows two users to produce a shared and symmetric key over an unsecured network.

According to Fig. 2 both parties, that is, client and server, agree on global elements prime P, and generator G which are both publicly available numbers; for instance, they may post the values on their web sites. They keep their respective private keys (X_A and X_G) secret, use modular division, and calculate respectively their shared (public) keys as Y_A and Y_B mathematically. The shared keys can take

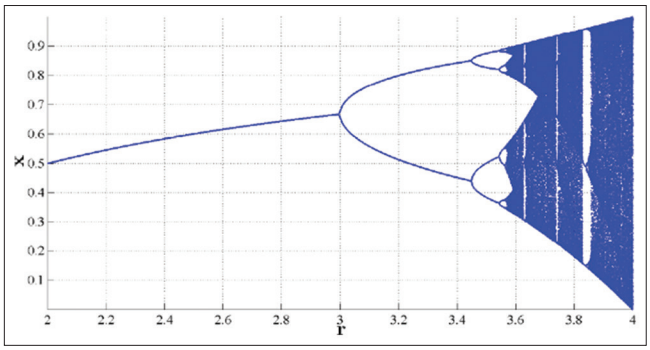


Fig. 1. The bifurcation schema of the 1D logistic map [13].

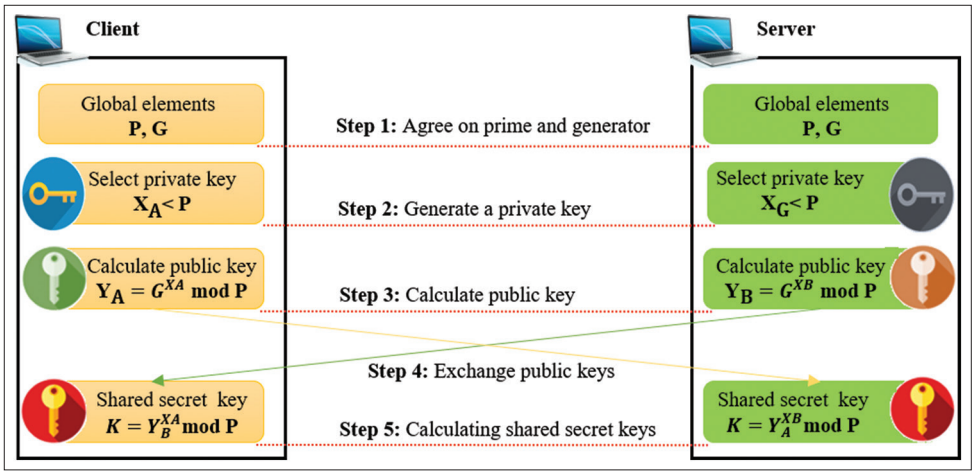


Fig. 2. Diffie–Hellman key exchange authentication procedure [15].

any value between 1 and (P – 1). Both parties exchange their shared keys and calculate the common secret keys value to be used in encrypting the image data which are transmitted over network. DHKE depends on a simple property of modular exponentiations. The mechanism of Diffie–Hellman key that exchanges mathematically is described as follows [15]–[17].

$$\left(G^{XA}\right)m\ o\ d\ P = \left(G^{XB}\right)m\ o\ d\ P \tag{3}$$

Where G, X, A, and P are positive integers. With existing $Y_A = G^A \bmod P$ and $Y_B = G^B \bmod P$, the value of k can be calculated without revealing A or B, which are called secret exponents.

$$K = \left(G^{XB}\right)m\ o\ d\ P \tag{4}$$

In this work, Diffie–Hellman key exchange mechanism was performed to generate and distribute the shared secret key K between the client and server parts. Then, the Diffie–Hellman encryption key K was defined and coded as a 256-bit integer number. In such a way, the encryption key was made to control the pseudo random sequences from the 2D logistic map algorithm for each round.

2.3. Medical Image Encryption Schemes

Protecting patient confidentiality and medical archives are an authorized requirement. Conventional cryptographic approaches are incapable of dealing with the massive quantity of medical image data and its appropriate statistical attributes. Numerous encryption mechanisms for securing medical privacy based on their content have been reported in the related literature. A modern two-dimensional Sine Logistic modulation map (2D-SLMM) which was derived from the logistic and sine maps was provided Hua *et al.* [18] with a broader chaotic spectrum than the traditional sine and chaotic logistic maps. In addition, they combined 2D-SLMM with a chaotic magic transform to establish a new image encryption algorithm (CMT). Simulation results and security analysis show that the proposed algorithm was able to protect images with minimal time complexity and a high level of security while also resisting numerous attacks. In Hua *et al.* [19], the authors addressed a 2D sine chaotification scheme to enhance the complex behavior of current chaotic maps. The authors applied their technique to improve the Henon map and 2D sinus logistics map. A generic medical image encryption system based on a new arrangement of two very powerful concepts, dynamic substitution boxes and chaotic maps, was introduced by Ibrahim *et al.* [20]. Before and after chaotic substitution, the arrangement of S-box substitution was seen to effectively avoid selected plaintext and cipher text attacks.

A two-dimensional of Sine Chaotification Model (2D-SCM) was recommended by Jo *et al.* [21]. The proposed technique can not only considerably improve the complexity of 2D chaotic maps but it can also significantly broaden their chaotic ranges. In Liu *et al.* [22], the authors recommended and implemented an Indistinguishability Under Chosen-Plaintext Attack (IND-CPA) secure CBIR architecture that executed image retrieval on the cloud without the constant interaction of the user. The author addressed a secure CBIR framework based on an Encrypted Difference Histogram (EDH-CBIR) in Liu *et al.* [23]. In that article, the image owner calculated the RGB component order or disorder difference matrices and encrypted them using

value replacement and position scrambling. The encrypted images were subsequently sent to the cloud server, which generated image feature vectors from encrypted difference histograms. To find similar images, image users encrypted the query image in the same way as the image owner did, and the cloud server extracted the query feature vector. To determine the degree of similarity, the Euclidean distance between the query feature vector and the image feature vector was determined.

In Lu *et al.* [24], the authors reported a secure CBIR framework that enables comparison of similarities between encrypted image features, depending on which the secure image retrieval based on content can be achieved. In the system, they focused on security strategies for image features that make a comparison of similarities between protected features. The authors in Sibahee *et al.* [25] developed an effective a lightweight system for content-based browsing over an encrypted image dataset using a Locality Sensitive Hashing (LSH) technique. The LSH index increased the system's expertise and effectiveness, enabling only relevant images to be obtained with a minimum of distance assessments. Vector refining methods were performed for efficient and safe refining of relevant results. The index building process ensured privacy of saved data and trapping doors. In 2017, the author of Gaata and Hantoosh [26] recommended an encrypted CBIR framework with additional improvement for the image retrieval based on features analysis. In this work, the gray level co-occurrence matrix with Haralick features, in combination with color moments, was used to construct the vector feature. In addition, Bloom filter and hash-table for image dataset classification were utilize. A similarity search procedure, based on secure transformation over encrypted cloud images, was presented in Xia *et al.* [27]. In the system, the authorized image client extracted and encrypted the feature vector with a query image to create the cipher text using the secure transformation method. The cipher text was then sent to the cloud to estimate the similarities of the vectors of the transformed feature.

The authors, in Bhagat and Gite [28], published an article on image retrieval for improved authentication using sparse codewords with cryptography. To prevent some form of attack, the Square Quality Control (SQC) encryption algorithm was utilized to the image. Lima *et al.* [29] employed a Cosine Number Transform (CNT) medical image encryption framework for the chosen Galois field. Huang *et al.* [30] implemented an encryption method, which combines chaos and Deoxyribonucleic Acid techniques. To achieve the permutation, substitution, and diffusion needed for encryption, they have performed a scenario of two rounds with six stages. The relatively low encryption speed was the key drawback of their system. The authors, in Hua *et al.* [31], proposed the Medical Image Encryption: Bitwise XOR (MIE-BX), high-speed scraping, random-pixel input, and pixel adaptive diffusion medical image encryption procedure. Their experimental outcomes indicated that the speed of their encryption method can outperform conventional approaches of encryption, such as Advanced Encryption Standard. However, Chen *et al.* [32] pointed out a serious vulnerability in MIE-BX to a reset attack known as the reset attack on Pseudo-Random Number Generator (PRNG). In this attack, the adversary restored the PRNG state to produce the same number of sequences each time.

The authors, in Badr *et al.* [3], introduced the dual authentication mechanism to support an attribute-based encryption (ABE) system that involves multiple parties such as data owners, data users, cloud servers, and authority. The proposed technique provided a safe solution to the problem of sharing significant information such as medical images. In Li *et al.* [33], a healthcare method was introduced by the authors as a new solution to the secure exchange of electronic health information between various entities in separate clouds. The proposed procedure was based on a revocable key policy and ABE algorithms that allowed clients to share encrypted health information on the basis of the data proprietors' and patients' own policy. The research paper in Wu *et al.* [13] applied the 2D logistic map of complicated basin structures and attractors for image encryption. In cryptography, the presented method implemented the classic permutation-substitution network structure; consequently, it ensures both uncertainty and diffusion properties for a protected cipher. In this work, the medical image encryption and decryption were constructed with the proposed system based on a two-dimensional (2D) chaotic map with Diffie–Hellman key exchange protocols to further enhance the security of the proposed system.

3. PROPOSED METHODOLOGY

3.1. System Architecture

In the proposed methodology section, the architecture overview is described, as depicted in Fig. 3. In the proposed framework, encryption and decryption techniques were applied as a force for securing medical images from unauthorized access. A content-based query operation was

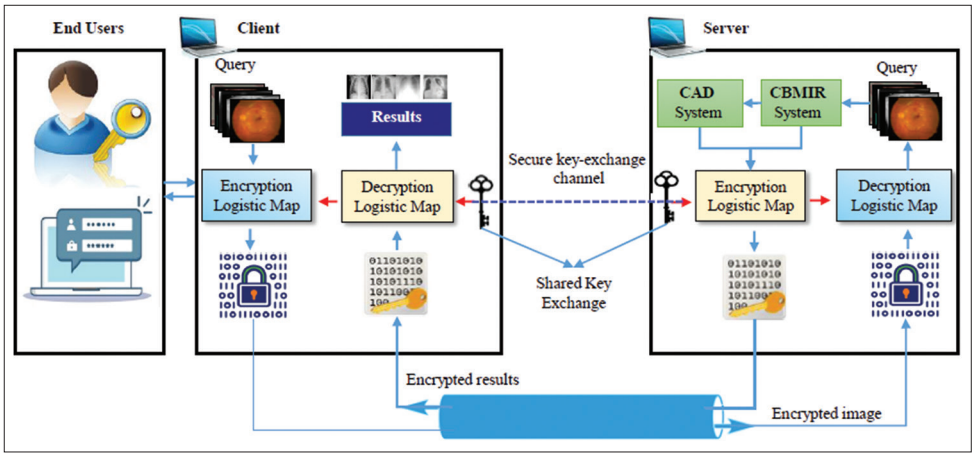


Fig. 3. Architecture overview of the proposed system operation scenario.

initiated from a client side to retrieve the similar medical images and diagnosis result with respect to a query image. Once the query image was chosen as input, 2D logistic map-based Diffie–Hellman key exchange protocol was applied to perform the encryption process for selected query image. After medical images were encrypted, their owner could upload them to the server for conducting the retrieval and classification tasks.

3.2. Encryption and Uploading Stage

First, the authorized users can login using their credentials to login to the proposed system. During this first authentication stage, both credentials are examined by checking the pre-existing dataset. Failure to enter credentials would obstruct the processing of users into additional steps. Now that, a user has registered, and they can start on the client side. On the client side, the query image is selected from a set of test images to carry out the retrieval and diagnosis of medical images. The pseudocode of the encryption procedure using 2D logistic map is shown in Algorithm 1 and the encryption process in the proposed schema involves the following stages:

1. Registering authorized users to access client-side portal
2. Browsing for the query image from the directory of images
3. Calculating the value of the shared secret key (K) between the Client-Server using Diffie–Hellman key exchange protocol
4. Encrypting the query image using the 2D logistic map algorithm with generated Diffie–Hellman key exchange key exchange secret key (Fig. 4).
5. Uploading the encrypted query image data to the server with the shared secret keys (K) value.

Algorithm 1: Pseudocode of image encryption

Input: Plain medical image Pi
Output: Cipher medical image Ci
Step 1: Parameter initialization
Read the original image and convert into a grayscale image
Use the Diffie–Hellman key exchange to generate the secret key (K)
Set Xo, Yo, r, T, F, S, and N
Translate K to map formats
Step 2: Image cipher (encryption)
For i = 1: N
Generate chaotic sequence using 2D logistic map
(e.g., Execute Equation 1)
Compute the 2D logistic permutation (pixel shuffling)
Apply the 2D logistic diffusion (pixel shifting)
Perform the 2D logistic transposition
End
Step 3: Produce cipher image

3.3. Downloading and Decryption Stage

After the client was authenticated by the server with an approved public key, the encrypted data, including the shared secret keys (K) value and medical query image, had to be downloaded and securely stored to the server. On the server side, the encrypted medical query image was directly decrypted using 2D logistic map algorithm and the same shared secret keys (K) from the key that was used for encryption process. The pseudocode of the decryption procedure using 2D logistic map is shown in Algorithm 2. Then, the decrypted image as shown in Fig. 5 was used as a query image to feed into content-based medical image retrieval (CBMIR) and CAD systems for retrieving and classification process of medical images. Following this, the retrieval and diagnosis results were encrypted and transmitted to the client using same procedure. Finally, the client received the returned encrypted results from the server side and performed the decryption process to get the final retrieval

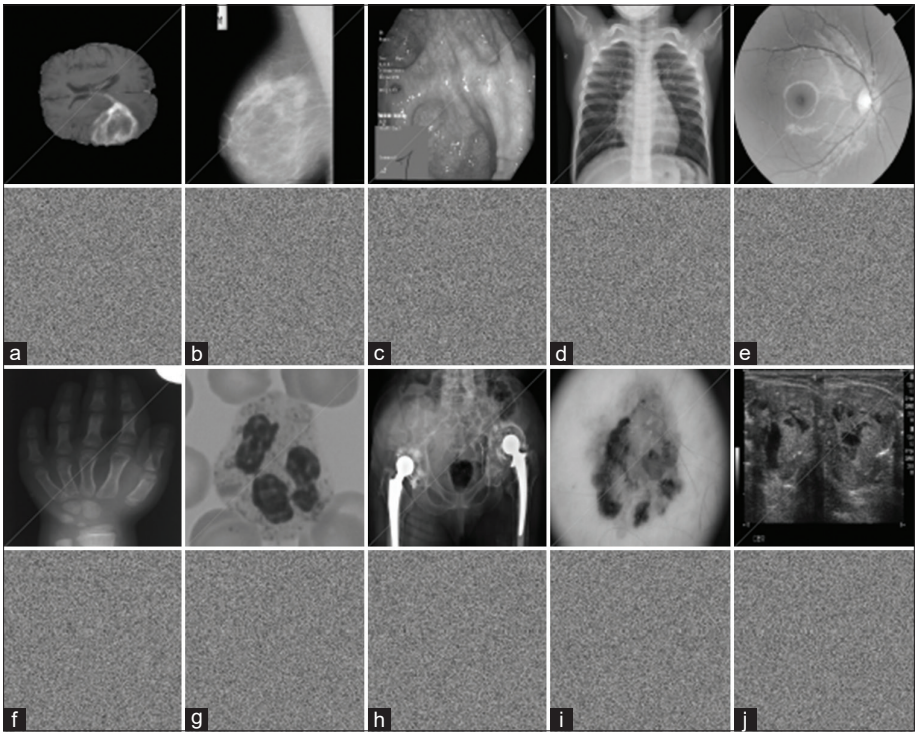


Fig. 4. Medical image encryption using 2D logistic map and Diffie–Hellman key exchange algorithms. original and encrypted medical image: (a) brain tumor; (b) breast mammogram; (c) cecum; (d) chest X-ray; (e) eye fundus; (f) hand X-ray; (g) leukemia; (h) pelvis X-ray; (i) clinical skin; and (j) thyroid images.

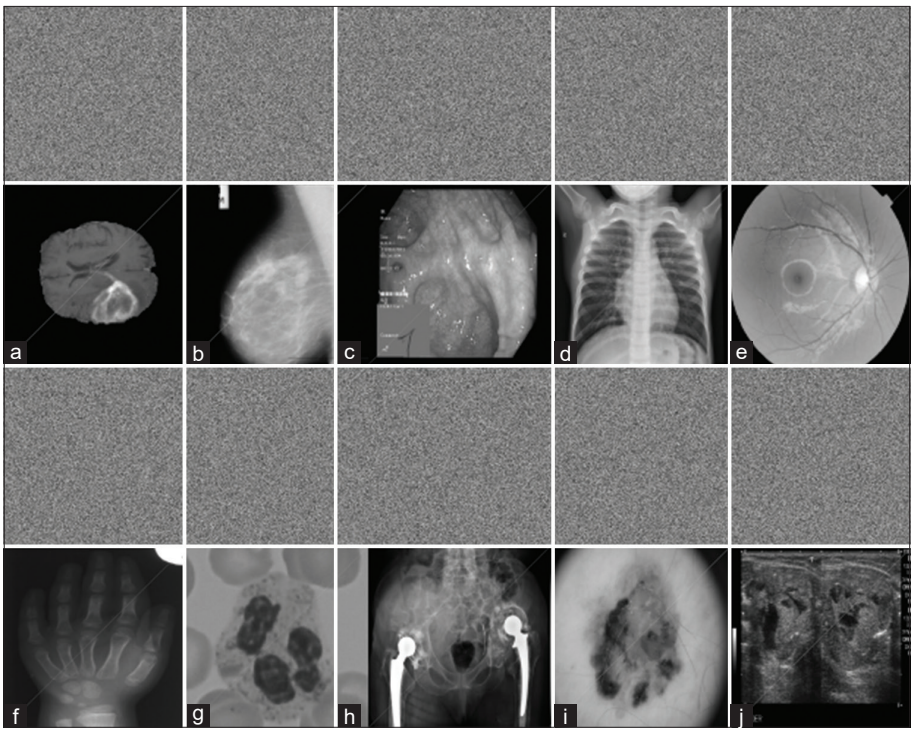


Fig. 5. Medical image decryption using 2D logistic map and Diffie–Hellman key exchange algorithms. encrypted and decrypted medical images: (a) brain tumor; (b) breast mammogram; (c) cecum; (d) chest X-ray; (e) eye fundus; (f) hand X-ray; (g) leukemia; (h) pelvis X-ray; (i) clinical skin; and (j) thyroid images.

and diagnosis results. The GUI of the client side is shown in Fig. 6.

Algorithm 2: Pseudocode of image decryption

Input: Cipher medical image Ci
Output: Decrypted medical image Di
Step 1: Parameter initialization
Read the original image and convert into a grayscale image
Use the Diffie–Hellman key exchange to generate the secret key (K)
Set Xo, Yo, r, T, F, S, and N
Translate K to map formats
Step 2: Image Cipher (Encryption)
For i = 1: N
Generate chaotic sequence using 2D logistic map (e.g., Execute Equation 1)
Perform the 2D logistic transposition
Apply the 2D logistic diffusion (pixel shifting)
Compute the 2D logistic permutation (pixel shuffling)
End
Step 3: Produce decrypted image

3.4. Objective Evaluation

Peak Signal-to-Noise Ratio (PSNR) and Mean Square Error (MSE) are two frequently used evaluation indicators. PSNR calculates the ratio between the maximum signal strength and the noise or distortion strength that affects the quality of the representation. In this paper, PSNR is calculated for the recovered (decrypted) image and original host image. The ratio between these two images is calculated in decibels (dB), where I and K denotes the original host image and recovered (decrypted) image. The PSNR is calculated according to the following equation:

$$PSNR\left(x,y\right)=10\log_{10}\left[\frac{\left[I\right]^2}{MSE}\right] \quad (5)$$

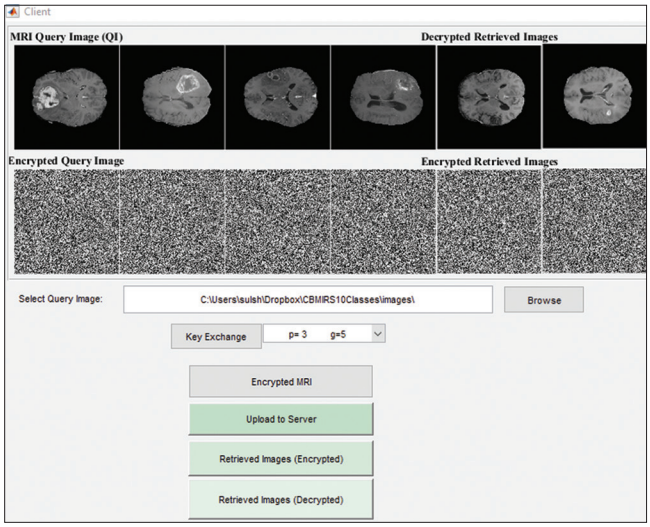


Fig. 6. The GUI of the client-side screen.

where *I* represent the maximum possible value of the pixel in the image (e.g., for a gray-scale image, the maximum value is 255). *MSE* calculates the magnitude of average error between the original image and recovered (decrypted) image. The *MSE* is computed as depicted below:

$$MSE\left(I,K\right)=\frac{1}{m\ n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left(\left[I\left(i,j\right)-\left|K\left(i,j\right)\right|\right]\right)^2 \quad (6)$$

Where:

- $I(i,j)$ denotes the pixel value in the coordinates (i, j) in the image *I* as a reference (original image)
- $K(i,j)$ denotes the pixel value at the coordinates (i, j) of the image *K* as being compared (decrypted image)
- m denotes the image height (in pixels)
- n denotes the image width (in pixels).

4. RESULTS AND ANALYSIS

A secure system for medical image retrieval and diagnosis is the big challenging task in the field of medical image processing. This work proposed a secure framework based on 2D logistic map algorithm along with Diffie–Hellman key exchange scheme for assuring the privacy and security of patient information. In this study, the proposed architecture was implemented in MATLAB 2019 (a) on MacBook Pro machine equipped with a processor of i7 (2.7 GHz Intel Core and 8 GB RAM). A fundamental prerequisite for an effective encryption scheme to avoid statistical attacks is the homogeneity of an encrypted image histogram [34], [35]. The distribution of input images can be represented by image histograms. To gain valuable information regarding the original image, an attacker may analyze the histogram of an encrypted image using attack procedures and statistical analysis of the encrypted image. It is important to guarantee that there are no statistical similarities between both the original image and the encrypted image. The histogram analysis clarifies the distribution of the pixels on an image by plotting the number of pixels at each degree of intensity. A visual analysis of the proposed scheme can be done observing the histograms of the medical images before and after the encryption. For visual inspection, Figs. 7-11 show sample histograms of the encrypted and decrypted medical images generated by the proposed framework for each of the evaluated query images. The histogram of the original medical images (plaintext images) demonstrates how the number of pixels at each gray level is graphically distributed.

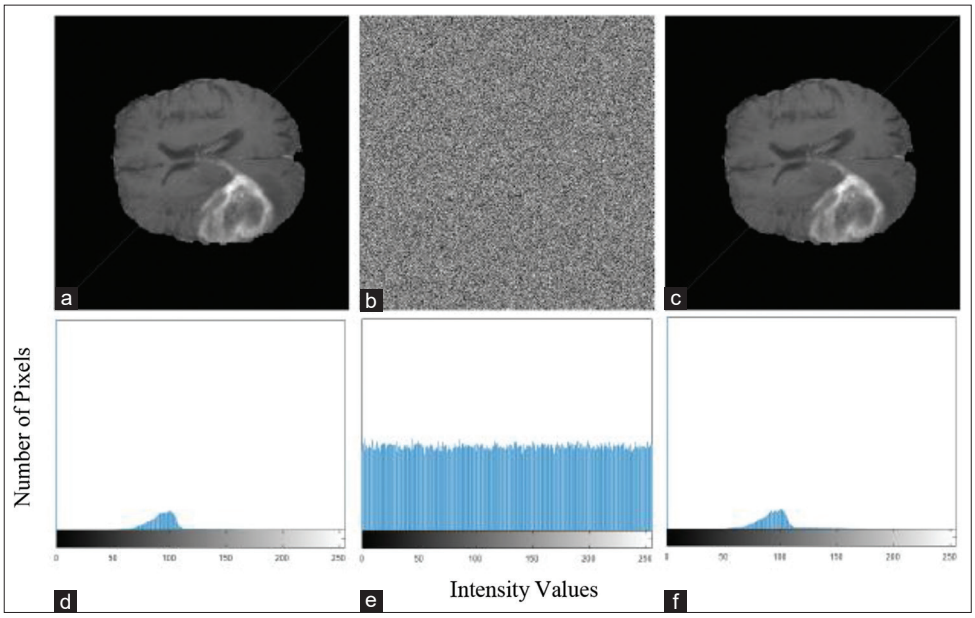


Fig. 7. Two-dimensional logistic map results with Diffie–Hellman on brain tumor image. (a) plaintext; (b) encrypted image; (c) decrypted image; (d) plaintext histogram; (e) encrypted histogram; and (f) decrypted histogram.

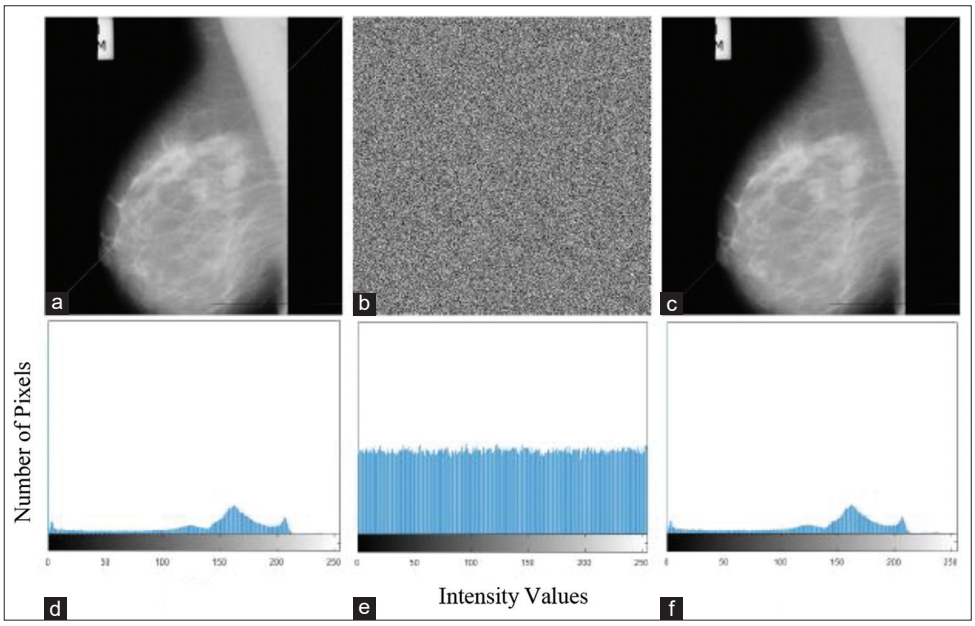


Fig. 8. Two-dimensional logistic map results with Diffie–Hellman on breast mammogram image. (a) plaintext; (b) encrypted image; (c) decrypted image; (d) plaintext histogram; (e) encrypted histogram; and (f) decrypted histogram.

It is evident that the encrypted image (ciphertext images) histogram is almost uniformly distributed, and that it obviously differs from the respective histograms in the original images. The procedure makes statistical attacks difficult, confirming that this evaluation is satisfied by the proposed framework. Therefore, in any statistical attack on the encryption of an image using the proposed technique, the encrypted image does not provide any proof for using. For the purpose of comparison of the recommended procedure, a sample of each different modality of medical (Magnetic Resonance, X-ray, Fundus, Endoscopy, and visible light) images was experienced in the tests. To evaluate the encryption

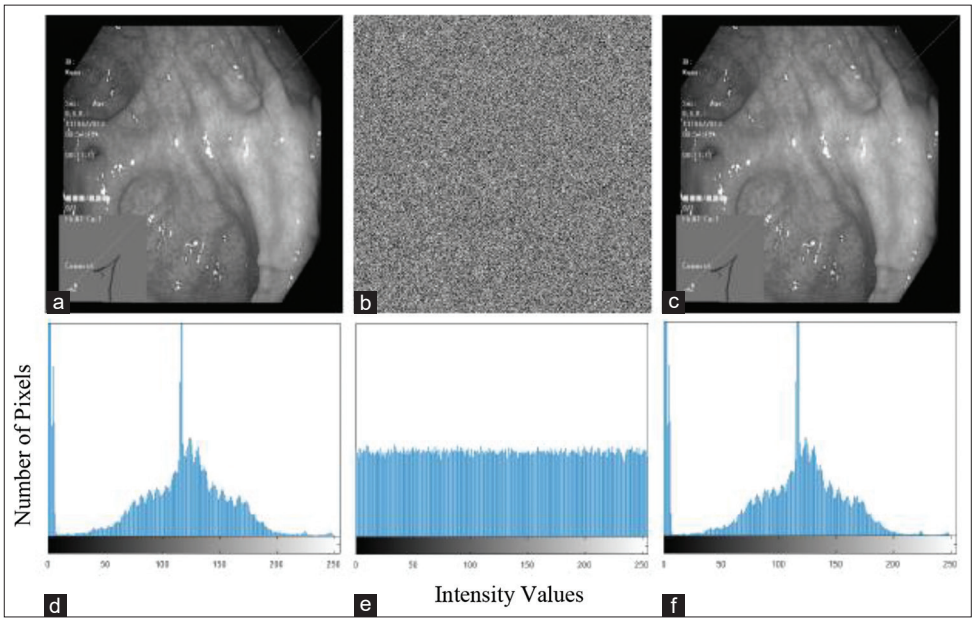


Fig. 9. Two-dimensional logistic map results with Diffie–Hellman on cecum image. (a) plaintext; (b) encrypted image; (c) decrypted image; (d) plaintext histogram; (e) encrypted histogram; and (f) decrypted histogram.

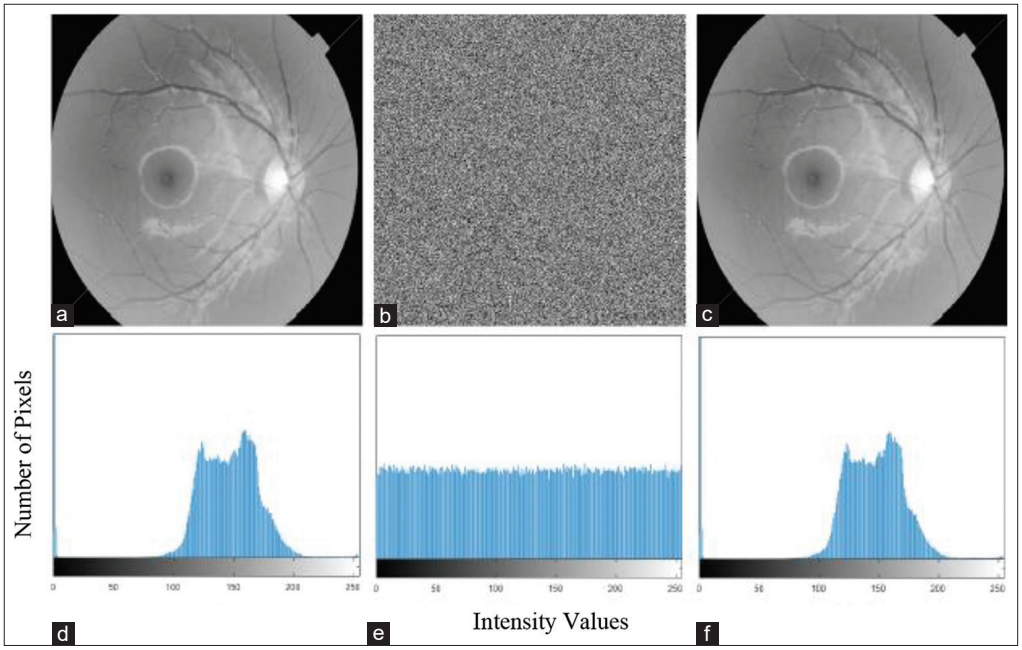


Fig. 10. Two-dimensional logistic map results with Diffie–Hellman on eye fundus image. (a) plaintext; (b) encrypted image; (c) decrypted image; (d) plaintext histogram; (e) encrypted histogram; and (f) decrypted histogram.

performance for different medical image modalities, the originality of the images was compared through their PSNR values. There will be no loss of data in the process which is an additional advantage of the proposed framework. Moreover, the proposed technique results in excellent performance with PSNR of 100% (a PSNR of 100 denotes no significant noise detected between the two images) and an MSE of zero (0) between encrypted and decrypted medical images.

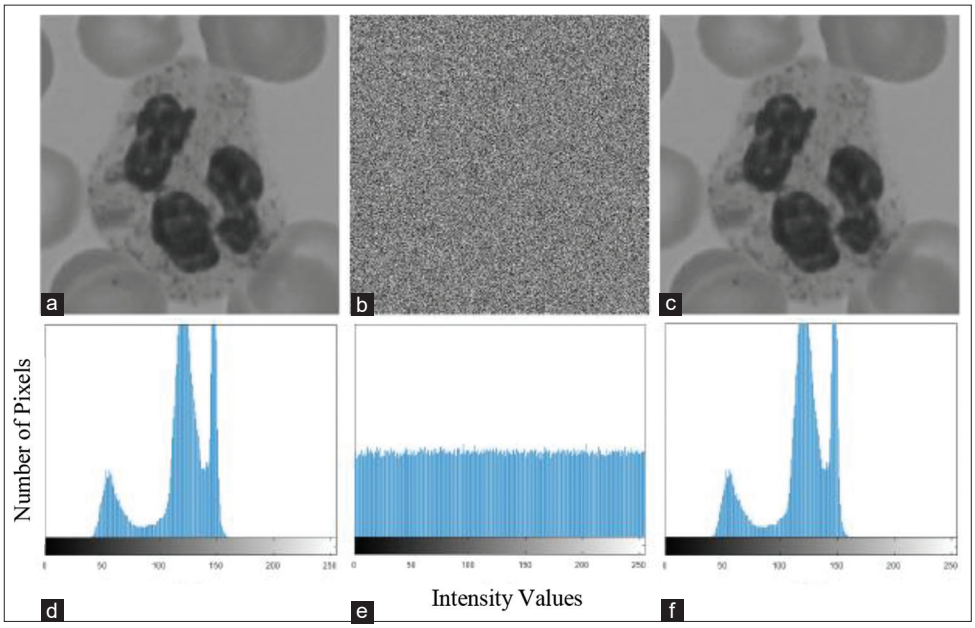


Fig. 11. Two-dimensional logistic map results with Diffie–Hellman on leukemia image. (a) plaintext; (b) encrypted image; (c) decrypted image; (d) plaintext histogram; (e) encrypted histogram; and (f) decrypted histogram.

5. CONCLUSION

Developing a secure medical image retrieval scheme has become an essential requirement for protecting the privacy of patients' medical information. This paper investigated a contemporary mechanism for employing client server to encrypt and exchange critical medical data, helping the user to avoid security risks. To achieve these security requirements, the proposed methodology used different procedures. To further improve the security of the proposed framework, medical image encryption and decryption were constructed based on a two-dimensional (2D) chaotic map with Diffie–Hellman key exchange protocols. The PSNR values were used to compare the image's originality. A further benefit of the proposed approach was that no data would be lost during the process. In addition, the presented approach generated outstanding results, with a PSNR of 100% (no substantial noise between the two images) and an MSE of zero (0) between encrypted and decrypted medical images.

REFERENCES

[1] S. Cheng, L. Wang and A. Du. "Histopathological image retrieval based on asymmetric residual Hash and DNA coding". *IEEE Access*, vol. 7, pp. 101388-101400, 2019.

[2] Z. Xia, L. Lu, T. Qiu, H. J. Shim, X. Chen and B. Jeon. "A privacy-preserving image retrieval based on ac-coefficients and color histograms in cloud environment". *Computers, Materials and Continua*, vol. 58, no. 1, pp. 27-43, 2019.

[3] A. M. Badr, Y. Zhang and H. G. A. Umar. "Dual authentication-based encryption with a delegation system to protect medical data in cloud computing". *Electronics*, vol. 8, no. 2, pp. 171, 2019.

[4] S. M. Farooq, S. M. S. Hussain, S. Kiran and T. S. Ustun. "Certificate based authentication mechanism for PMU communication networks based on IEC 61850-90-5". *Electronics*, vol. 7, p. 370, 2018.

[5] J. Fridrich. *Secure Image Ciphering Based on Chaos*. Final Report for AFRL, New York, 1997.

[6] J. Fridrich. "Symmetric ciphers based on two-dimensional chaotic maps". *International Journal of Bifurcation and Chaos*, vol. 8, no. 6, pp. 1259-1284, 2011.

[7] Z. Hua, Y. Zhou and H. Huang. "Cosine-transform-based chaotic system for image encryption". *Information Sciences*, vol. 480, pp. 403-419, 2019.

[8] J. A. P. Artiles, D. P. B. Chaves and C. "Pimentel. Image encryption using block cipher and chaotic sequences". *Signal Process Image Communication*, vol. 79, pp. 24-31, 2019.

[9] H. Zhu, Y. Zhao and Y. Song. "2D logistic-modulated-sine-coupling-logistic chaotic map for image encryption". *IEEE Access*, vol. 7, pp. 14081-14098, 2019.

[10] R. Guesmi, M. A. B. Farah, A. Kachouri and M. Samet. "A novel chaos-based image encryption using DNA sequence operation and secure hash algorithm SHA-2". *Nonlinear Dynamics*, vol. 83, no. 3, pp. 1123-1136, 2015.

[11] M. B. Yassein, S. Aljawarneh, E. Qawasmeh, W. Mardini and Y. Khamayseh. "Comprehensive study of symmetric key and asymmetric key encryption algorithms". In: *Proceeding 2017 International Conference on Engineering and Technology ICET*, pp. 1-7, 2017.

[12] R. Hamza, K. Muhammad, A. Kumar and G. "Ramirez-Gonzalez. Hash based encryption for keyframes of diagnostic hysteroscopy". *IEEE Access*, vol. 6, pp. 60160-60170, 2018.

[13] Y. Wu, J. P. Noonan, G. Yang and H. Jin. “Image encryption using the two-dimensional logistic chaotic map”. *Journal of Electronic Imaging*, vol. 21, no. 1, p. 3014, 2012.

[14] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg and N. Khandelwal. “Content-based image retrieval system for pulmonary nodules: Assisting radiologists in self-learning and diagnosis of lung cancer”. *Journal of Digital Imaging*, vol. 30, no. 1, pp. 63-77, 2017.

[15] S. M. Farooq, S. M. S. Hussain, S. Kiran and T. S. Ustun. “Certificate based authentication mechanism for PMU communication networks based on IEC 61850-90-5”. *Electronics*, vol. 7, no. 12, p. 370, 2018.

[16] A. Chopra. “Comparative analysis of key exchange algorithms in cryptography and its implementation”. *IMS Manthan* (*The Journal Innovations*), vol. 8, no. 2, 2015.

[17] H. Bodur and R. Kara. “Implementing diffie-hellman key exchange method on logical key hierarchy for secure broadcast transmission”. In: *Proceeding 9th International Conference on Computational Intelligence Communications Networks*, *CICN*, pp. 144-147, 2018.

[18] Z. Hua, Y. Zhou, C. M. Pun and C. L. P. Chen. “2D sine logistic modulation map for image encryption”. *Information Sciences*, vol. 297, pp. 80-94, 2015.

[19] Z. Hua, Y. Zhou and B. Bao. “Two-dimensional sine chaotification system with hardware implementation”. *IEEE Transactions on Industrial Informatics*, vol. 16, pp. 887-897, 2020.

[20] S. Ibrahim, H. Alhumyani, M. Masud, S. S. Alshamrani, O. Cheikhrouhou, G. Muhammad and A. M. Abbas. “Framework for efficient medical image encryption using dynamic S-boxes and chaotic maps”. *IEEE Access*, vol. 8, pp. 160433-160449, 2020.

[21] H. J. Jo, S. J. Gotts, R. C. Reynolds, P. A. Bandettini, A. Martin, R. W. Cox and Z. S. Saad. “Effective preprocessing procedures virtually eliminate distance-dependent motion artifacts in resting state FMRI”. *Journal of Applied Mathematics*, vol. 2013, pp. 935154, 2013.

[22] F. Liu, Y. Wang, F. C. Wang, Y. Z. Zhang and J. Lin. “Intelligent and secure content-based image retrieval for mobile users”. *IEEE Access*, vol. 7, pp. 119209-119222, 2019.

[23] D. Liu, J. Shen, Z. Xia and X. Sun. A content-based image retrieval scheme using an encrypted difference histogram in cloud computing. *Informatics*, vol. 8 no. 3, pp. 96, 2017.

[24] W. Lu, A. Varna, A. Swaminathan and M. Wu. “Secure image retrieval through feature protection”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. ICASSP, Taiwan, 2009.

[25] M. A. Al Sibahee, S. Lu, Z. A. Abduljabbar, A. Ibrahim, Z. A. Hussien, K. A. A. Mutlaq and M. A. Hussain. “Efficient encrypted image retrieval in IoT-cloud with multi-user authentication”. *International Journal of Distributed Sensor Networks*, vol. 14, 2018.

[26] M. T. Gaata MT and F. F. Hantoosh. “Encrypted image retrieval system based on features analysis”. *Al-Mustansiriyah Journal of Sciences*, vol. 28, no. 3, pp. 166-173, 2017.

[27] Z. Xia, Y. Zhu, X. Sun and J. Wang. “A similarity search scheme over encrypted cloud images based on secure transformation”. *International Journal of Future Generation Communication and Networking*, vol. 6, no. 6, pp. 71-80, 2013.

[28] M. N. Bhagat and P. B. B. Gite. “Image retrieval using sparse codewords with cryptography for enhanced security”. *IOSR Journal of Computer Engineering*, vol. 16, no. 2, pp. 22-26, 2014.

[29] J. B. Lima, F. Madeiro and F. J. R. Sales. “Encryption of medical images based on the cosine number transform”. *Signal Process Image Communication*, vol. 35, pp. 1-8, 2015.

[30] L. Huang, S. Wang, J. Xiang and Y. Sun. “Chaotic color image encryption scheme using deoxyribonucleic acid (DNA) coding calculations and arithmetic over the galois field”. *Mathematical Problems in Engineering*, vol. 2020, pp. 1-2, 2022.

[31] Z. Hua, S. Yi and Y. Zhou. “Medical image encryption using high-speed scrambling and pixel adaptive diffusion”. *Signal Processing*, vol. 144, pp. 134-144, 2018.

[32] Y. Chen, C. Tang and R. Ye. “Cryptanalysis and improvement of medical image encryption using high-speed scrambling and pixel adaptive diffusion”. *Signal Processing*, vol. 167, pp. 107286, 2020.

[34] M. Li, S. Yu, Y. Zheng, K. Ren and W. Lou. “Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption”. *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, pp. 131-143, 2013.

[35] H. Liu, B. Zhao, J. Zou, L. Huang, Y. Liu. “A lightweight image encryption algorithm based on message passing and chaotic map”. *Security and Communication Networks*, vol. 2004, no. 4, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Malicious URL Detection Using Decision Tree-based Lexical Features Selection and Multilayer Perceptron Model

Warmn Faiq Ahmed, Noor Ghazi M. Jameel

Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq

ABSTRACT

Network information security risks multiply and become more dangerous. Hackers today generally target end-to-end technology and take advantage of human weaknesses. Furthermore, hackers take advantage of technology weaknesses by applying various methods to attack. Nowadays, one of the greatest dangers to the modern digital world is malicious URLs, and stopping them is one of the biggest challenges in the field of cyber security. Detecting harmful URLs using machine learning and deep learning algorithms have been the subject of various academic papers. However, time and accuracy are the two biggest challenges of these tools. This paper proposes a multilayer perceptron (MLP) model that utilizes two significant aspects to make it more practical, lightweight, and fast: Using only lexical features and a decision tree (DT) algorithm to select the best relevant subset of features. The effectiveness of the experimental outcomes is evaluated in terms of time, accuracy, and error reduction. The results show that a MLP model using 35 features could achieve an accuracy of 94.51% utilizing only URL lexical features. Furthermore, the model is improved in time after applying the DT as feature selection with a slight improvement in accuracy and loss.

Index Terms: Multilayer Perceptron, Lexical Feature, Feature Selection, Malicious URL, Synthetic Minority Oversampling Technique

1. INTRODUCTION

The internet expands at an unprecedented rate. Most of the time, malicious software is spread via the internet. Malicious websites can be referred to as any website that has been designed to cause harm. It is similar to a legitimate URL for regular users but hosts unsolicited content. The attacker usually builds a website identical to the target or embeds the exploit code of browser vulnerabilities on the webpage. Then, it tricks the victim into clicking on these links to obtain the victim's information or control the victim's computer [1]. In many circumstances, people do not check the complete website URL, and the attacker can obtain essential and personal information once they visit a malicious website [2].

Malicious URL detection always comes at the top in the research area. However, having protection against these attacks is not an option anymore. According to Google's Transparency Report, 2.195 million websites made their list of "Sites Deemed Dangerous by Safe Browsing" category as of January 17, 2021. The vast majority of those (over 2.1 million) were phishing sites. Only 27,000 of Google's removed websites were delisted because of malware [3]. Several forms of a malicious URL proceed with the attack and deliver unsolicited content, mainly named spam, phishing, and drive-by download. Spam is a web page with many links to unwanted websites for other purposes; the

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp105-116

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: warmn.faiq.a@spu.edu.iq

Received: 20-08-2022 Accepted: 01-10-2022 Published: 13-11-2022

pages may pretend to provide assistance or facts about a subject. Phishing is a type of social engineering attack used to steal sensitive data. Finally, drive-by downloads refer to the unintentional download of malicious code to the device, leaving it open to a cyber-attack [4].

There are currently several approaches to detect dangerous websites on the internet. Nowadays, a malicious URL is mainly detected by black and white list-based and machine learning-based URL detection methods. According to the first technique, a website cannot be viewed until the URL is checked against the blacklist database to ensure it is not on the list. Blacklist is essentially a listing of URLs that were previously identified as malicious. Its advantage is that it is fast, easy, and has a meager false-positive (FP) rate. However, the main problem with this method is that it has a high false-negative (FN) rate and fails to detect newly generated URLs [1], [5], [6]. Nevertheless, it has been widely utilized in several major browsers, including Mozilla Firefox, Safari, and Chrome, among others, due to its simplicity and efficiency [5]. In addition, the blacklisting approach is also utilized by many antivirus systems and internet businesses. However, due to some limitations, the blacklisting strategy is insufficient to identify non-blacklisted threats [7]. Whitelist is another aspect that provides security when accessing a website. It is similar to the blacklist method technique. The difference is that in the whitelist, only those websites are allowed to access that is in the list. The limitation of this method is denying access to many newly generated websites that are legal and safe to visit [5]. On the other hand, machine learning techniques use a collection of URLs specified as a set of attributes and train a prediction model based on them to categorize a URL as good or bad, enabling them to recognize new, possibly harmful URLs [1].

In this paper, the multilayer perceptron (MLP) model is used to detect malicious URLs based on the features of the URLs. Since a lightweight method is challenging for time efficiency, lexical features are utilized and extracted from the dataset to train the model. The model is tested first without and then with feature selection (FS) to see the result and the differences. The main contribution of this paper is the development of a malicious URL detection system that utilizes only lexical features to construct a light model and selects only high-ranked features to reduce feature extraction (FE) time. Moreover, using decision tree (DT) as a FS algorithm is an advantage to select the best relevant features based on features importance score to improve the model performance and decrease the FE time during the detection process.

The paper is organized as follows. Section 2 is related works. The proposed malicious URL detection system with its phases including dataset collection, features extraction, features selection using DT algorithm, model development, and evaluation is presented in Section 3. All the experimental results and discussions are provided in Section 4. Finally, Section 5 illustrates the conclusion of the paper.

2. RELATED WORKS

Many kinds of research in the area of detecting malicious websites with various techniques, algorithms, and methods exist. The machine learning technique is one of the approaches used to solve the problem of malicious URL detection. Multiple studies have been done in the era. Xuan *et al.* proposed support vector machine (SVM) and random forest (RF) as machine learning algorithms to classify benign and malicious URLs by extracting features and behaviors of the URLs. The researchers created an extensive set of features to improve the model's ability and use it as a free tool to detect malicious URLs [8]. Subha *et al.* tested various machine learning algorithms to detect malicious URLs. According to the results, RF scored better than all SVM, Naïve Base, and artificial neural network (ANN) with an accuracy of 97.98 and the F1 score of 92.88 [9]. Furthermore, Islam *et al.* used three machine learning algorithms to detect malicious URLs: NN, K-nearest neighbor (KNN), DT, and RF. The results showed that the neural network (NN) scored the worst, whereas DT and RF achieved the best scores. The study mentioned that the lack of ability to detect malicious URLs by NN is due to the small size of the dataset, while NN is suitable for large datasets [10].

Besides, some of the researches used NNs as a solution for classifying malicious URLs from benign ones. Liu and Lee proposed a detection method using a convolutional neural network (CNN). The research adopted the end user's perspective and used CNN to learn and recognize screenshot images of the websites. The results showed that although the training period is lengthy, it is tolerable, especially with powerful graphics processing units. The testing is efficient once the training is completed; therefore, time is often not an issue with this procedure [11]. Balamurugan *et al.* proposed a NN to classify the websites as good and bad URLs with optimizing network parameters using genetic algorithms. The article showed a good improvement when optimizers were applied to the NN model in both classification and convergence [12]. Furthermore, Chen *et al.* used CNN for malicious URL detection. The study showed that the

proposed method achieved satisfying detection accuracy with an accuracy of 81.18% [13].

Moreover, hybrid systems are also proposed by some recent studies as a solution to the problem. Naresh *et al.* proposed a machine learning-based system that combines a SVM with logistic regression using a combination of URL lexical options, payload size, and python supply options as features to recognize the malicious URLs. As a result, an accuracy of 98% was achieved, which is an improvement compared to a conventional method. According to some recent articles, using NNs as a hybrid system can achieve satisfying performance [14]. Yang *et al.* proposed a system to detect malicious websites based on integrated CNNs and RF system. The results showed that the proposed integrated system achieved better results than traditional machine learning algorithms due to their shallow design, which cannot examine the complicated link between safe and malicious URLs [2]. Another research is by Das *et al.* who tested three NN algorithms, RNN, LSTM, and CNN-LSTM, to see the effectiveness of these algorithms in classifying benign and malicious URLs. The results showed that with an accuracy of 93.59%, the CNN-LSTM architecture exceeds the other two [15]. Furthermore, Peng *et al.* proposed attention-based CNN-LSTM for malicious URL detection. The results showed that the proposed method achieved better than shallow NNs and single deep NNs such as CNN and LSTM individuals with an accuracy of 96.74 [16].

3. THE PROPOSED MALICIOUS URL DETECTION SYSTEM

The proposed system is constructed using a lightweight method. Only lexical features are utilized to build the model. Python is used for programming the phases of the proposed system with famously fast and reliable libraries such as Pandas, Numpy, Scikit-learn, Imblearn, Pyplot, TensorFlow, and Keras.

The architecture of the proposed system starts with loading the dataset and then preprocessing stages to prepare the data for training. The training stage starts after the data are prepared. Then the testing stage; the trained model classifies whether the URL is malicious or benign. Finally, evaluation metrics are applied to compute the performance of the model. The system architecture is shown in Fig. 1.

3.1. Dataset Collection

In this work, a proposed model was trained and tested on a dataset conducted from malicious and benign websites that were utilized to create the suggested model and evaluate its predictions [17]. The dataset initially consisted of 420,464 URLs, 344,821 benign (good), and the rest of 75,643 websites are malicious (bad), as shown in Table 1. Therefore, the number of URLs in each class is imbalance, as shown in Fig. 2. A sample of the instances is shown in Fig. 3.

3.2. Data Preprocessing

3.2.1. Data cleaning

One of the most critical preprocessing stages in machine learning is data cleaning. Having clean, accurate noiseless data give precise models and results. Starting with cleaning the data, 9216 duplicated URLs were found and removed. The dataset was then checked for missing values, and there were no missing values in the dataset.

3.2.2. URL Lexical Feature Extraction

Several characteristics separate a safe URL and its webpage from a malicious URL. In certain instances, attackers employ direct IP linkages rather than domain names. Another tactic use by attackers is short names or abbreviations for websites unrelated to legitimate brand names. Algorithms for the detection method involve a wide variety of characteristics. To detect malicious websites using machine learning techniques, several distinct characteristics were retrieved from various academic research, such as lexical, host-based, and content-based features.

Since lexical features are fast to extract, they are also more applicable due to facing some casual problems when using content-based and host-based features. Most of the time, content-based features cannot be extracted from malicious URLs since most are blacklisted and cannot be accessed to get the contents such as HTML, JavaScript, and visual features. Besides, the security risks when accessing such websites need precautions such as using special sandbox services to reduce the risk. Host-based FE also faces problems such as a very long time taking due to the vast number of online requests from the database servers such as WHOIS that sometimes lead to another problem: Closing sockets for some of the websites and not getting the required information. In this study, lexical features are utilized to recognize malicious websites and distinguish them from legitimate ones. These characteristics are derived from the URL address's elements like a string. It should be able to identify malicious URLs because it bases its decision on how the URL appears. By replicating the names and making minor modifications, many attackers may make dangerous URLs seem normal. However, from the perspective of machine learning, it is not feasible to take the actual name of the URL. Instead, the URL's string must be handled to obtain valuable properties. Sixty lexical

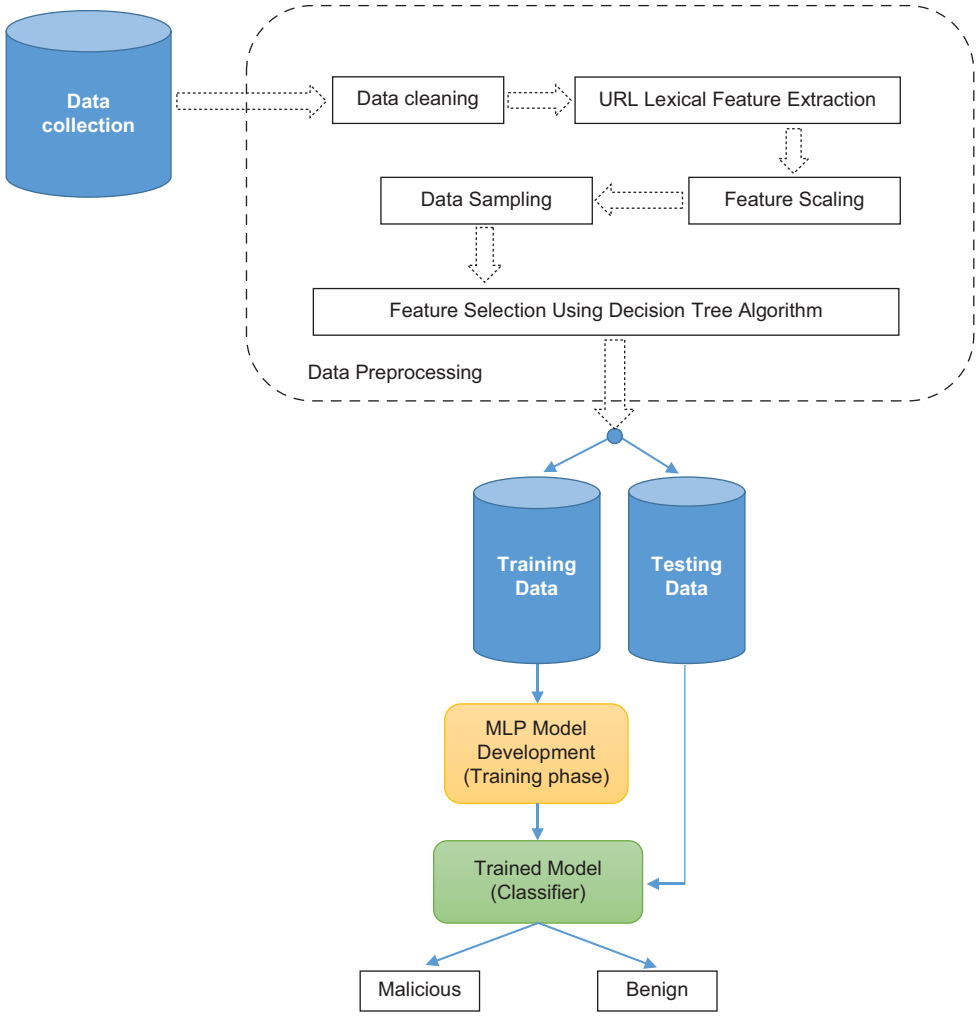


Fig. 1. The proposed system architecture.

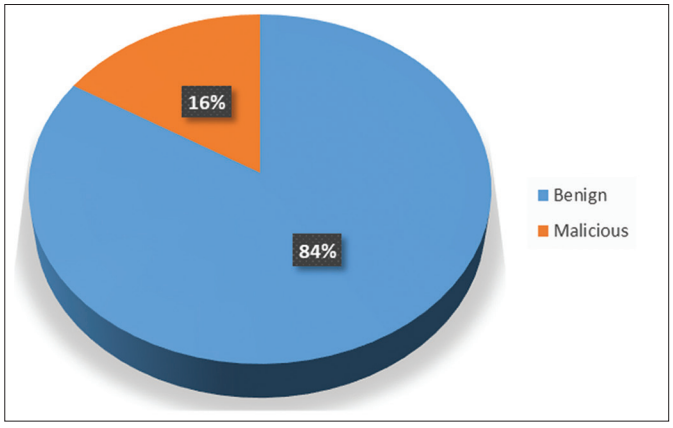


Fig. 2. Dataset class distribution.

features were collected from literature, then extracted from the web links as listed in Table 2.

TABLE 1: Dataset description

Type	No. of URLs
Benign	344,821
Malicious	75,643
Total URLs	420,464

3.2.4. Feature scaling

Feature scaling or normalization is often advised and sometimes crucial. Normalization is vital for NNs since unnormalized inputs to activation functions might cause trapping in a relatively flat domain region. Feature scaling helps optimize NN algorithms by accelerating training and preventing optimization from being trapped in local optima. Models of NNs establish a mapping between input and output variables. As a result, each variable's size and distribution of the data extracted from the domain may change. Input variables can have distinct scales because of

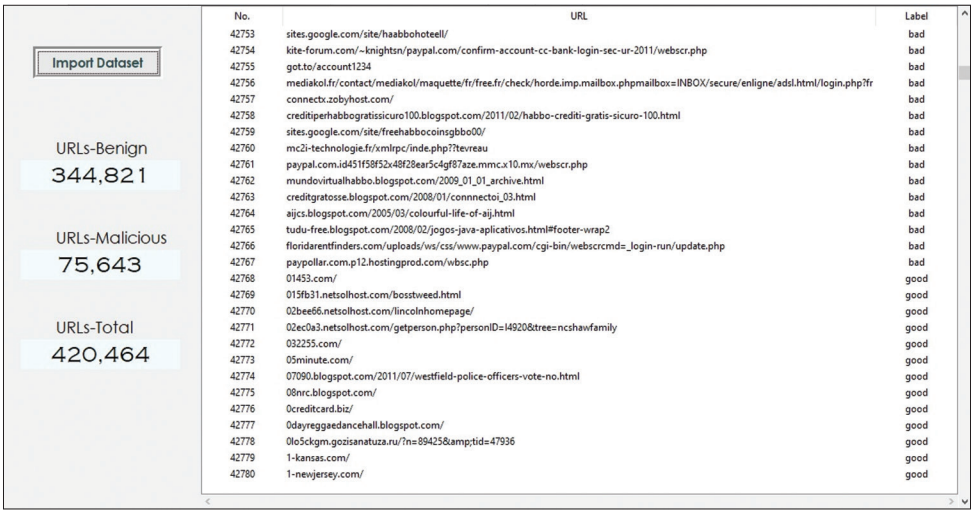


Fig. 3. Sample of the dataset instances.

TABLE 2: List of URL lexical features

Feature No.	Feature names	Data type	Description	References
f0	Count dots	Integer	Number of character “.” in URL	[7], [8], [18]-[21]
f1	url depth	Integer	The depth of the URL	[8]
f2	url length	Integer	The length of the URL	[7], [8], [14], [16], [18]-[20], [22]-[26]
f3	hyphen	Integer	Number of the dash character “-” (hyphen)	[8], [20], [22], [23]
f4	AT symbol	Boolean	There exists a character “@” in URL	[8], [22], [23], [27]
f5	Tide symbol	Boolean	There exists a character “~” in URL	[8]
f6	numUnderscore	Integer	Number of the underscore character	[8], [22]
f7	numPercent	Integer	Number of the character “%”	[8], [20]
f8	numAmpersand	Integer	Number of the character “&”	[8], [20], [22]
f9	numHash	Integer	Number of the character “#”	[8], [22]
f10	countQuestionMark	Integer	count the number of “?” in url	[20]
f11	countSemicolon	Integer	count the number of “;” in URL	[22]
f12	httpsInUrl	Boolean	Check if there exists a HTTPS in website URL	[8], [19], [22], [28]
f13	ipAddress	Boolean	Check if the IP address is used in the hostname of the website URL	[7], [8], [16], [22], [23], [25]
f14	urlRedirection	Boolean	There exists a slash “//” in the link path	[8], [19], [22], [23], [27]
f15	Count alpha	Integer	Number of the alphabetic character	[20], [22]
f16	Alpha ratio	Floating point	The proportion of alphabetic characters in the URL to the total length of the URL	[22]
f17	Count digit	Integer	Number of the numeric character	[8], [20], [22], [29]
f18	Digit ratio	Floating point	The proportion of numeric characters in the URL to the total length of the URL	[22]
f19	Count special chars	Integer	Number of any special characters like”,' %”,”$”,”,’ =”, etc.	[4], [7], [8], [14], [16], [18], [19], [22], [24]-[26]
f20	Special chars ratio	Floating point	The proportion of special characters in the URL to the total length of the URL	[16], [22]
f21	Count lowercase	Integer	The number of lowercase English letters in the URL	[16], [22]
f22	Lowercase ratio	Floating point	The proportion of lowercase English letters in the URL to the total length of the URL	[16], [22]
f23	Count uppercase	Integer	The number of uppercase English letters in the URL	[16], [22]
f24	Uppercase ratio	Floating point	The proportion of uppercase English letters in the URL to the total length of the URL	[16], [22]
f25	Count_subdomain	Integer	Number of subdomains in the URL	[8], [18]
f26	Short URL	Boolean	Using tiny url/short url service	[14], [23], [25]
f27	Length_of_ hostname	Integer	Length of hostname	[8], [19]

(Contd...)

TABLE 2: *(Continued)*

Feature No.	Feature names	Data type	Description	References
f28	Length_of_path	Integer	Length of the link path	[8], [19], [20]
f29	Length_of_query	Integer	Length of the query	[8], [20]
f30	Length_of_scheme	Integer	Length of the URL scheme	[20]
f31	Presence_sus_file_ext	Boolean	Checking the URL string for the presence of the following file extensions·exe,·scr,·vbs,·js,.xml, .docm,·xps, .iso, .img, doc, .rtf,·xls, pdf, .pub, .arj, .lzh, .r01, .r14, .r18, .r25, .tar, .ace, .zip, .jar, .bat, .cmd, .moz, .vb, .vbs, .js, .wsc, .wsh, .ps1, .ps1×ml, .ps2, .ps2×ml, .psc1 and .psc2.	[25]
f32	Count_ar_num	Integer	The number of Arabic numerals in the URL	[16]
f33	Is_tld_in_top5	Boolean	Whether the top-level domain is the top five domains (com, cn, net, org, cc)	[16]
f34	Paypal_in_path	Boolean	If “paypal” is contained in the PATH section.	[30]
f35	Ali_in_path	Boolean	If “ali” is contained in the PATH section.	[30]
f36	Jd_in_path	Boolean	If “jd” is contained in the PATH section.	[30]
f37	Safety_in_path	Boolean	If “safety” is contained in the PATH section.	[30]
f38	Verify_in_path	Boolean	If “verify” is contained in the PATH section.	[30]
f39	Google_in_path	Boolean	If “Google” is contained in the PATH section.	[30]
f40	Apple_in_path	Boolean	If “apple” is contained in the PATH section. if_facebook_u	[30]
f41	Facebook_in_path	Boolean	If “Facebook” is contained in the PATH section.	[30]
f42	Amazon_in_path	Boolean	If “amazon” is contained in the PATH section.	[30]
f43	Porn_in_path	Boolean	If “porn”-related words are contained in the PATH section.	[30]
f44	Gamble_in_path	Boolean	If “gamble” related words are contained in the PATH section.	[30]
f45	Paypal_in_domain	Boolean	If “paypal” is contained in the DOMAIN section.	[30]
f46	Ali_in_domain	Boolean	If “ali” is contained in the DOMAIN section.	[30]
f47	Jd_in_domain	Boolean	If “jd” is contained in the DOMAIN section.	[30]
f48	Safety_in_domain	Boolean	If “safety” is contained in the DOMAIN section.	[30]
f49	Verify_in_domain	Boolean	If “verify” is contained in the DOMAIN section.	[30]
f50	Google_in_domain	Boolean	If “Google” is contained in the DOMAIN section.	[30]
f51	Apple_in_domain	Boolean	If “apple” is contained in the DOMAIN section.	[30]
f52	Facebook_in_domain	Boolean	If “Facebook” is contained in the DOMAIN section.	[30]
f53	Amazon_in_domain	Boolean	If “amazon” is contained in the DOMAIN section.	[30]
f54	Porn_in_domain	Boolean	If “porn” related words are contained in the DOMAIN section.	[30]
f55	Gamble_in_domain	Boolean	If “gamble” related words are contained in the DOMAIN section.	[30]
f56	Has keyword “client”	Boolean	If the word “client” is contained in the URL	[31]
f57	Has keyword “admin”	Boolean	If the word “admin” is contained in the URL	[31]
f58	Has keyword “server”	Boolean	If the word “server” is contained in the URL	[31]
f59	Has keyword “login”	Boolean	If the word “login” is contained in the URL	[31]

their varied. The difficulty of the problem being modeled could be exacerbated by differences in the scales across the input variables. A model may learn tremendous weight values due to large input values, such as a spread of thousands of units, makes the result to be biased toward the bigger units. When features are of comparable size and nearly normally distributed, several machine learning methods work better or converge more quickly. Min-max algorithm is used to scale all the features between 0 and 1. Equation (1) uses for min-max feature scaling which helps the model to understand and learn better and faster without biasing to the more significant values [20].

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where, x_{max} and x_{min} are the maximum and the minimum values of the feature (x), respectively.

3.2.5. Data sampling

Initial examination of the dataset revealed that there were 5.18 times fewer occurrences of harmful websites than benign ones. Therefore, due to the stark disparity in the number of malicious and benign website instances, the model affect to be biased due to this significant class imbalance

as it learns from a far higher percentage of benign website occurrences.

A balanced class dataset is necessary for classification issues. As most machine learning algorithms used for classification were developed based on the presumption that there are an equal number of instances of each class, the imbalance of types in classification presents problems for predictive modeling. Therefore, a balanced classification dataset is also necessary for a classification model to produce accurate judgments.

There are several ways to handle an imbalanced dataset. The synthetic minority oversampling technique (SMOTE) was utilized to address this issue. The SMOTE technique uses KNN machine learning algorithm to produce new instances. Using it, additional instances of the minority class have been created, matching the proportion of instances of each class to the majority class to balance the classes. To balance the dataset, the minority class must thus be oversampled unless both groups have almost an equal number of cases. After balancing, the minority class were oversampled, which caused the data size to grow. Finally, the 344,800 occurrences of each class result in a balanced distribution, as shown in Fig. 4.

3.2.6. Feature Selection using DT Algorithm

The quality of FS and importance is one of the crucial differentiators in every machine learning task. Due to computational limitations and the need to remove noisy variables for more accurate prediction, FS becomes necessary when there is a large amount of data that the model may process.

In this study, a DT algorithm is used to select the best and most relevant lexical features based on the feature importance score. DTs apply various techniques to decide whether to divide a node into two or more sub-nodes. The homogeneity of newly formed sub-nodes is increased by sub-node formation. The threshold value of an attribute is used to divide the nodes in the DT into sub-nodes. The classification and regression tree algorithm uses the Gini index criteria to find the sub-nodes with the best homogeneity. The DT divides the nodes based on all factors that are accessible before choosing the split that produces the most homogenous sub-nodes. At the same time, the target variables are considered while selecting an algorithm. It is a visual depiction of every option for making a choice based on specific criteria according to the algorithm. Conditions on any characteristics are used to make judgments in both situations. The leaf nodes reflect the selection based on the conditions, whereas the inside nodes represent the conditions. Finding the attribute that provides the most information is necessary for DT construction. By building the tree in this way, feature importance scores can be accessed and used to help interpret the data, ranking, and select features that are most useful to a predictive model. It aids in determining which variable is chosen to be used in producing the decisive internal node at a specific point. The steps of FS using a DT are described in an (Algorithm 1). At this phase, the list of features with their importance values is calculated and selected by the DT algorithm.

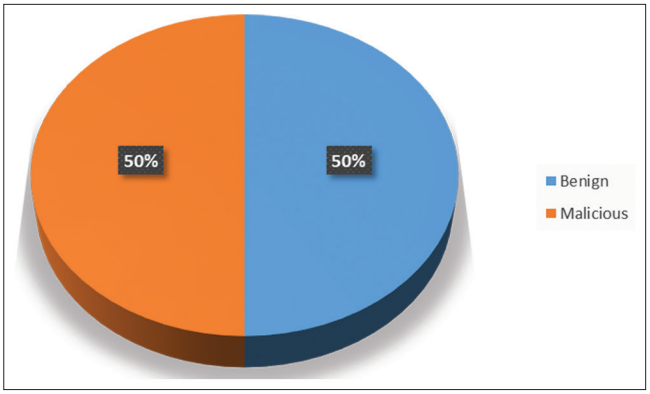


Fig. 4. Dataset after data sampling using SMOTE.

Begin
repeat
Step-1: Start at the Root node with all instances *S*
Step-2: Select an attribute based on splitting criteria
minGini ← 0
splitTree ← ∅
for all attributes a in S **do**
$giniIndex \leftarrow 1 - \sum_a P_a^2$
if giniIndex < minGini **then**
minGini ← giniIndex
splitA ← a
end if
end for
Step-3: Partition instances according to selected attribute recursively
CART(attributes-a, instances, target attribute)
until all instances processed
Step-4: There are no remaining attributes for further partitioning
End

Algorithm 1. Classification and regression tree [32].

3.3. MLP Model

The most practical variety of NNs is MLP which is frequently used to refer to the area of ANNs. A perceptron is a single-neuron model that serves as the basis for more extensive NNs. Artificial neurons are the basic units of NNs. The feed-

TABLE 3: The parameters of the proposed MLP model

Layer no.	No. of neurons/dim	Optimizer	Activation function	Learning rate	Batch size	No. of epochs
Layer 1	400	Adam	Sigmoid	0.005	200	1500
Layer 2	300		Sigmoid			
Layer 3	200		Sigmoid			

forward NN is supplemented by the MLP. There are three layers: The input layer, the output layer, and the hidden layer.

The proposed MLP model consists of three hidden layers besides the input and output layers to describe the model. The first hidden layer has 400 neurons, the second hidden layer has 300 neurons, and the last hidden layer has 200 neurons. The output layer has one neuron as it is a binary classification with two outputs, 1 and 0, whereas 1 represents a malicious URL and 0 represents a benign one. The other parameters are set as a batch size of 200, a learning rate of 0.005, a sigmoid function as an activation function, and Adam as an optimizer, as shown in Table 3.

3.4. Model Evaluation

The goal is not just to create a predictive model. It involves building and choosing a model that performs well on out-of-sample data. Therefore, verifying the model's correctness is essential before computing estimated values. To assess the models, many indicators are considered. A crucial phase in the machine learning pipeline is evaluating the learned model's effectiveness. Machine learning models are either adaptable or non-adaptive based on how effectively they generalize to new input. When an ML model is applied to new data without being adequately evaluated using a variety of metrics and without relying on accuracy, it may produce inaccurate predictions. Besides, the accuracy, precision, recall, and F1 score have been taken into account for the model reliability and considering the aspect of the errors when the model classifies between malicious and benign URLs. The definition of classification accuracy, which may be the most straightforward criterion to use and apply, is the ratio of correct predictions to all other predictions and calculated using Equation (2) [33].

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \quad (2)$$

Confusion matrix produces a matrix that summarizes the overall effectiveness of the model. For example, the confusion matrix for binary classification, which is the case in this work, is a two-by-two matrix. The confusion matrix shows the number of correct and incorrect classification for both actual and predicted values, including true positive indicates the

TABLE 4: Confusion matrix

Actual values	Predicted values	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

TP: True positive, TN: True negative, FP: False positive, FN: False negative

TABLE 5: List of used hardware and software specifications

Hardware and software specification	Description
PC	Core i3 gen6
RAM	20 GB
Storage	SSD SATA 256 GB
Operation system	Windows 10 pro

number of samples that are correctly classified as positive and true negative shows the number of instances that are correctly identified as negative, besides, there is FP that indicates the number of samples that are incorrectly identified as positive, and finally, FN that indicates the number of instances that are incorrectly identified as negative. The confusion matrix for binary classification is shown in Table 4.

From the confusion matrix, some important metrics are calculated and taken into consideration along with the accuracy to ensure that the model performs well and is not biased because of issues such as dataset imbalance. Therefore, precision, recall, and F1 score are used as model evaluation metrics. Precision indicates how accurate the positive predictions are, recall is the coverage of actual positive samples, and the F1 score is the harmonic mean of precision and recall, and they are calculated using Equations (3), (4), and (5), respectively [22], [29], [34].

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

TABLE 6: List of features with their importance score

Feature No.	Feature importance	Feature No.	Feature importance	Feature No.	Feature importance	Feature No.	Feature importance
f0	0.11828	f16	0.05732	f32	0	f48	0
f1	0.07532	f17	0.04211	f33	0.07169	f49	0
f2	0.03691	f18	0.04414	f34	0.00132	f50	0
f3	0.01727	f19	0.01206	f35	0.00158	f51	0
f4	0.00161	f20	0.13231	f36	0.00022	f52	0
f5	0.00185	f21	0.02187	f37	0.00009	f53	0
f6	0.01472	f22	0.02058	f38	0.00041	f54	0
f7	0.00227	f23	0.00755	f39	0.00241	f55	0
f8	0.0018	f24	0.01264	f40	0.00031	f56	0.00053
f9	0.00009	f25	0.02609	f41	0.00228	f57	0.01168
f10	0.00874	f26	0.01038	f42	0.00009	f58	0.00089
f11	0.00997	f27	0.1204	f43	0.00017	f59	0.02466
f12	0.00017	f28	0.05412	f44	0		
f13	0.00007	f29	0.00539	f45	0		
f14	0.00058	f30	0.00068	f46	0		
f15	0.01788	f31	0.0065	f47	0		

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the details of the experimental results are presented. The experiments are implemented on a malicious URL dataset [19] aiming to find the set of relevant URL lexical features based on their importance score using DT algorithm and evaluating the MLP model performance using the selected features. The final prepared dataset after the main steps of data preprocessing which includes data cleaning, data sampling, and FE, consists of a total of 689,600 URLs with 60 lexical features and a class label that has a 0 for benign and 1 for malicious. The software and hardware specifications used for the experiments are explained in Table 5.

After running the DT algorithm for FS, the importance score or weight for each variable was calculated. Features with lowest importance scores were deleted and features with highest scores were kept. This type of FS can simplify the problem that is being modeled, speed up the modeling process, and improve the performance of the model. The list of all lexical features' importance scores is illustrated in Table 6. After this phase, 35 features were selected and 25 features were eliminated. The selected features are the top 35 features with highest importance values which are f0, f1, f2, f3, f4, f5, f6, f7, f8, f10, f11, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f25, f26, f27, f28, f29, f31, f33, f34, f35, f39, f41, f57, f58, and f59.

As a result of eliminating 25 features, a significant decrease in FE time achieved, which is an essential factor in this problem situation, as shown in Table 7 and Fig. 5.

TABLE 7: Feature extraction time before and after feature selection

No. of features	Feature extraction time in seconds
60 features, the whole dataset (Before FS)	134 s
35 features, whole dataset (After FS)	92 s

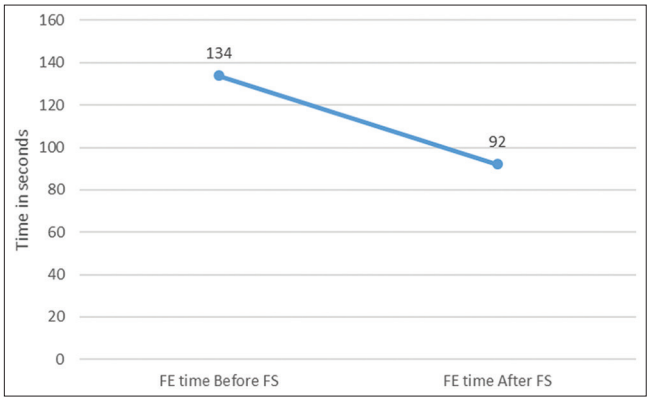


Fig. 5. FE time differences before and after FS.

For MLP model evaluation, the 35 selected features were fed to the model as input. The stratified technique was used for splitting the dataset into train and test sets to preserve the same proportions of instances in each class as in the original dataset. It is obvious that most of the data in the dataset are advised to be used for training to let the model learn well. Different ratios for training and testing have been used by the researchers such as 80% for training and the other 20% for testing or 70% for training by 30% for testing. Many factors are taken into consideration when train test split is done, such as the number of instances in the dataset, hyperparameters

TABLE 8: List of tested scenarios

Scenario	No. of epochs	No. of features	Batch size	Learning rate	No. of neurons in hidden layers
s1	100	35	200	0.005	200, 120, 80
s2	100	35	200	0.005	400, 200, 100
s3	100	35	200	0.005	400, 300, 200
s4	100	35	200	0.005	600, 400, 200
s5	100	35	200	0.005	800, 600, 400
s6	500	35	200	0.005	400, 300, 200
s7	500	35	200	0.005	600, 400, 200
s8	500	35	200	0.005	800, 600, 400
s9	1000	35	200	0.005	400, 300, 200
s10	1500	35	200	0.005	400, 300, 200

TABLE 9: Results of all the 10 scenarios

Scenarios	Train time in seconds	Test time in seconds	Train loss	Train accuracy (%)	Test accuracy (%)	Precision	Recall	F- score	Confusion matrix
s1	933.4	15.0	0.145	93.90	92.82	0.923	0.935	0.929	([95321 8119] [6735 96705])
s2	2258.5	28.7	0.142	94.00	92.95	0.919	0.943	0.930	([94797 8643] [5938 97502])
s3	2553.3	17.8	0.123	94.79	93.45	0.927	0.944	0.935	([95733 7707] [5840 97600])
s4	2847.4	23.3	0.122	94.86	93.51	0.927	0.944	0.936	([95807 7633] [5798 97642])
s5	6984.2	31.8	0.125	94.74	93.51	0.935	0.936	0.935	([96659 6781] [6636 96804])
s6	10487.2	18.3	0.091	96.21	94.18	0.937	0.948	0.942	([96822 6618] [5415 98025])
s7	17460.9	25.3	0.098	96.00	94.08	0.939	0.943	0.941	([97118 6322] [5918 97522])
s8	27800.3	37.7	0.095	96.09	94.15	0.937	0.946	0.942	([96877 6563] [5546 97894])
s9	22684.7	19.7	0.086	96.49	94.25	0.938	0.947	0.943	([97010 6430] [5460 97980])
s10	62791.6	30.3	0.075	96.93	94.51	0.941	0.950	0.945	([97233 6207] [5146 98294])

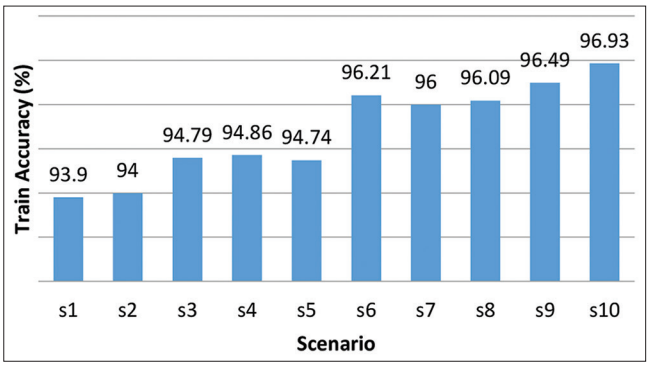


Fig. 6. Train accuracy for the 10 different scenarios.

to tune, the used classifier, and the model use case. Due to the good amount of instances in the dataset, 70% of the final dataset considered for training, while the remaining 30% is used for testing. The model with several scenarios has been tested using a learning rate of 0.005, batch size of 200, and different number of epochs and neurons. The list of scenarios is described in Table 8.

After executing all the 10 scenarios described in Table 8, from the results shown in Table 9, it is obvious that with increasing the number of epochs, the accuracy will increase along with training time, and the training loss will decrease eventually. In this system, the more important parameters for detecting malicious URLs are higher values for test accuracy, precision, and recall with lower training loss. The least important parameter is the training time. Training phase is a one-time process, sometimes it requires a long time to develop a well-trained model with high accuracy and less training loss. Since the last scenario, 1500 epochs outperformed the best scores for the mentioned parameters, it has been chosen to train the model and used for malicious URL detection. As a result,

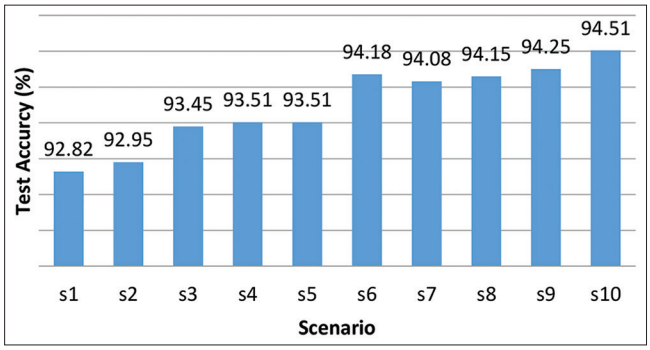


Fig. 7. Test accuracy for the 10 different scenarios.

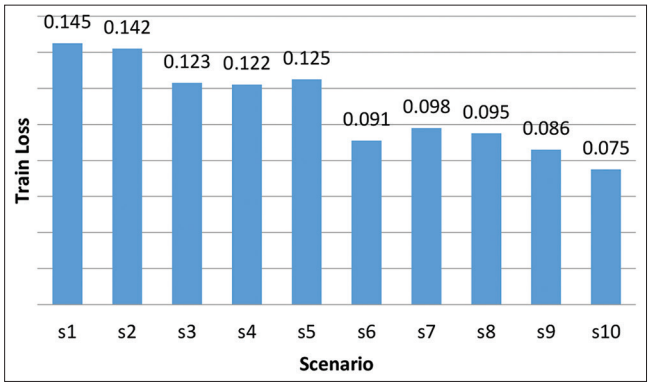


Fig. 8. Train loss for the 10 different scenarios.

the model achieved an accuracy of 94.51, recall of 94.1, the precision of 95.0, and training loss of 0.075. The results are shown in Table 9 and illustrated in Figs. 6-8.

5. CONCLUSION

One of the serious threats on the internet is malicious URL. Hackers have several techniques and algorithms to obfuscate URLs to bypass the defenses. The problem of detecting malicious URLs has been studied in this research with explaining types of possible attacks, features, and detection techniques. The study developed a lightweight malicious URL detection model using URL lexical features only instead of content or host-based features. Content and host-based features take a long time during the extraction. To extract content-based features, the websites should be available for accessing their source code. Host-based features extraction process needs connection with special servers such as WHOIS to get the required information. DT has been used to get the importance scores of all lexical features to select the best features to build a malicious URL detection system with better performance and efficiency. The study shows that using only relevant lexical features, which is more practical to apply, is enough to create a robust lightweight detection model using MLP algorithm. Experiment results have been shown and discussed to explain the differences before and after applying each technique.

REFERENCES

[1] J. Yuan, G. Chen, S. Tian and X. Pei. "Malicious URL detection based on a parallel neural joint model," *IEEE Access*, vol. 9, pp. 9464-9472, 2021.

[2] R. Yang, K. Zheng, B. Wu, C. Wu and X. Wang. "Phishing website detection based on deep convolutional neural network and random forest ensemble learning," *Sensors*, vol. 21, no. 24, pp, 8281, 2021.

[3] S. Cook. "Malware Statistics in 2022: Frequency, Impact, Cost and More," 2022. Available from: https://www.comparitech.com/antivirus/malware-statistics-facts [Last accessed on 2022 Aug 18].

[4] S. Kumi, C. Lim and S. G. Lee. "Malicious url detection based on associative classification." *Entropy*, vol. 23, no. 2, pp. 1-12, 2021.

[5] W. Bo, Z. B. Fang, L. X. Wei, Z. F. Cheng and Z. X. Hua. "Malicious URLs detection based on a novel optimization algorithm." *IEICE Transactions on Information and Systems*, vol. E104.D, no. 4, pp. 513-516, 2021.

[6] Z. Chen, Y. Liu, C. Chen, M. Lu and X. Zhang. "Malicious URL detection based on improved multilayer recurrent convolutional neural network model." *Security and Communication Networks*, vol. 2021, pp. 9994127, 2021.

[7] S. M. Nair. "Detecting malicious URL using machine learning: A Survey." *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 5, pp. 2670-2677, 2020.

[8] C. Do Xuan, H. Dinh Nguyen and T. Victor Nikolaevich. "Malicious URL Detection Based on Machine Learning." *International Journal of Advanced Computer Science and Applications*, vol. 11, pp. 148-153, 2020.

[9] V. Subha, M. S. Pretha and R. Manimegalai. " Malicious Url Classification Using Data Mining Techniques." *Journal of Analysis and Computation (JAC)*, pp. 148-153, 2018.

[10] M. Maminur Islam, S. Poudyal and K. Datta Gupta. "Map reduce implementation for malicious websites classification." *International Journal of Network Security and its Applications*, vol. 11, no. 5, pp. 27-35, 2019.

[11] D. Liu and J. H. Lee. "Cnn based malicious website detection by invalidating multiple web spams." *IEEE Access*, vol. 8, pp. 97258-97266, 2020.

[12] P. Balamurugan, T. Amudha, J. Satheeshkumar and M. Somam. "Optimizing neural network parameters for effective classification of benign and malicious websites." *Journal of Physics Conference Series*, vol. 1998, no. 1, 2021.

[13] Y. Chen, Y. Zhou, Q. Dong and Q. Li. "A Malicious URL detection method based on CNN." In: *2020 IEEE Conference on Telecommunications, Optics and Computer Science, TOCS 2020*. IEEE, Piscataway, 2020, pp. 23-28.

[14] N. Khan, R. Naresh, A. Gupta and S. Giri. "Ayon gupta and sanghamitra Giri, malicious URL detection system using combined SVM and logistic regression model." *International Journal of Advanced Research in Science, Engineering and Technology*, vol. 11, no. 4, pp. 63-73, 2020.

[15] A. Das, A. Das, A. Datta, S. Si and S. Barman. "Deep approaches on malicious URL classification." In: *2020 11th International*

Conference on Computer Networks and Communication Technologies. ICCCNT 2020, IEEE, Piscataway, 2020.

[16] Y. Peng, S. Tian, L. Yu, Y. Lv and R. Wang. "Malicious URL recognition and detection using attention-based CNN-LSTM." *KSII Transactions on Internet and Information Systems*, vol. 13, no. 11, pp. 5580-5593, 2019.

[17] Adamyong. "GitHub-Adamyong-zbf/URL_Detection: Data Set." 2020. Available from: https://github.com/adamyong-zbf/URL_detection [Last accessed on 2022 Aug 18].

[18] L. M. Camarinha-Matos, N. Farhadi, F. Lopes and H. Pereira, Editors., *Technological Innovation for Life Improvement*, Vol. 577. Springer International Publishing, Cham, 2020.

[19] S. Singhal, U. Chawla and R. Shorey. "Machine learning concept drift based approach for malicious website detection." In: *2020 International Conference on Communication Systems Networks*, *COMSNETS 2020*, IEEE, Piscataway, pp. 582-585, 2020.

[20] Maheshwari S, B. Janet and R. J. A. Kumar. "Malicious URL Detection: A Comparative Study." In: *Proceedings International Conference on Artificial Intelligence and Smart Systems*, *ICAIS 2021*. IEEE, Piscataway, pp. 1147-1151, 2021.

[21] Y. Peng, S. Tian, L. Yu, Y. Lv and R. Wang. "A Joint Approach to Detect Malicious URL Based on Attention Mechanism." *International Journal of Computational Intelligence and Applications*, vol. 18, no. 3, 2019.

[22] A. S. Raja, R. Vinodini and A. Kavitha. "Lexical features based malicious URL detection using machine learning techniques." *Materials Today Proceedings*, vol. 47, pp. 163-166, 2021.

[23] S. D. Vara Prasad and K. R. Rao. "A Novel Framework For Malicious URL Detection Using Hybrid Model." *Turkish Journal of Computer and Mathematics Education*, vol. 12, pp. 2542, 2021.

[24] S. Ahmad and A. Tamimi, "Detecting Malicious Websites Using Machine Learning," M.S. thesis, Department of Graduate Programs & Research, Rochester Institute of Technology, RIT Dubai, April. 2020. [Online]. Available from: https://scholarworks.rit.edu/theses

[25] T. Manyumwa, P. F. Chapita, H. Wu and S. Ji. "Towards Fighting Cybercrime: Malicious URL Attack Type Detection using Multiclass Classification." In: *Proceedings 2020 IEEE International Conference on Big Data, Big Data 2020*, IEEE, Piscataway, pp. 1813-1822, 2020.

[26] F. Alkhudair, M. Alassaf, R. Ullah Khan and S. Alfarraj. "*Detecting Malicious URL*." IEEE, Piscataway, 2020.

[27] R. R. Rout, G. Lingam and D. V. L. Somayajulu. "Detection of malicious social bots using learning automata with url features in twitter network." *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 1004-1018, 2020.

[28] Y. C. Chen, Y. W. Ma and J. L. Chen. "Intelligent malicious url detection with feature analysis." In: *Proceedings Second IEEE Symposium on Computer and Communications*. Vol. 2020. IEEE, Piscataway, 2020.

[29] S. He, J. Xin, H. Peng and E. Zhang. "Research on malicious URL detection based on feature contribution tendency." In: *2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics*, *ICCCBDA 2021*, pp. 576-581, 2021.

[30] T. Li, G. Kou and Y. Peng. "Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods." *Information Systems*, vol. 91, pp. 101494, 2020

[31] R. Ikwu. In: R. E. Ikwu, editor. "*Extracting Feature Vectors From URL Strings For Malicious URL Detection*." Towards Data Science," Canada, 2021. Available from: https://towardsdatascience.com/extracting-feature-vectors-from-url-strings-for-malicious-url-detection-cbafc24737a [Last accessed on 2022 Aug 16].

[32] G. S. Kori and D. M. S. Kakkasageri. "*Classification and Regression Tree (Cart) Based Resource Allocation Scheme for Wireless Sensor Networks*." Social Science Research Network, Rochester, NY, 2022.

[33] N. Hosseini, F. Fakhar, B. Kiani and S. Eslami. "Enhancing the security of patients' portals and websites by detecting malicious web crawlers using machine learning techniques." *International Journal of Medical Informatics*, vol. 132, pp. 103976, 2019.

[34] M. Chatterjee and A. S. Namin. "Deep Reinforcement Learning for Detecting Malicious Websites." *Computer Science*, vol. 15. pp. 55, 2019.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Kurdish Speech to Text Recognition System Based on Deep Convolutional-recurrent Neural Networks

Lana Sardar Hussein, Sozan Abdulla Mahmood

Department of Computer Science, College of Science, University Sulaimanyah, Sulaimanyah, Kurdistan Region, Iraq

ABSTRACT

In recent years, deep learning has had enormous success in speech recognition and natural language processing. In other languages, recent progress in speech recognition has been quite promising, but the Kurdish language has not seen comparable development. There are extremely few research papers on Kurdish speech recognition. In this paper, investigated Gated Recurrent Units (GRUs) which is one of the popular RNN models to recognize individual Kurdish words, and propose a very simplified deep-learning architecture to get more efficient and high accuracy model. The proposed model consists of a combination of CNN and GRU layers. The Kurdish Sorani Speech KSS dataset was created for the speech recognition system, as its 18799 sound files for 500 formal Kurdish words. Finally, the model proposed was trained with collected data and yielded over %96 accuracy. The combination of CNN an RNN (GURs) for speech recognition achieved superior performance compared to the other feed-forward deep neural network models and other statistical methods.

Index Terms: Deep Learning, Gated Recurrent Units, Kurdish Speech Recognition, Convolutional Neural Network

1. INTRODUCTION

Speech is a natural means for people to communicate with one another. Automatic Speech Recognition (ASR) is the technique through which a computer can recognize spoken words and understand what they are saying. The ASR is the first component of a smart system. It is a method of converting an auditory signal into a string of words, which can then be used as final outputs or inputs in natural language processing. The purpose of ASR systems is to recognize human-spoken natural languages. ASR technology is commonly utilized in computers with speech interfaces, foreign language applications, dictation, hands-free operations and controls, and other features that enable interactions between machines and humans faster and easier than using keyboards [1].

ASRs are designed using a variety of methodologies, the most notable of which is the Hidden Markov Model (HMM) and machine learning-based methods, such as Artificial Neural Networks (ANNs) and Convolutional Neural Network (CNN) [2]. Increasing the accuracy and efficiency of these systems is one of the issues that still exist in this sector. Deep learning, a relatively new technology, has been widely employed to address this issue. Because an audio signal is a sample of sequential data, meaning its present value is reliant on all past values, in this work RNN (GRU) applied in addition with CNN. RNN is a type of artificial neural network. It involves a sequential data connection with the hidden neurons. It can be applied for the applications of text, audio, and video. It deals with sequential data from the

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp117-125

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Lana Sardar Hussein, Department of Computer Science, College of Science, University of Sulaimanyah, Sulaimanyah, Kurdistan Region, Iraq. lana.salih@univsul.edu.iq

Received: 29-04-2022 Accepted: 07-09-2022 Published: 18-11-2022

analyzed the sequence at each time depending on the previous time in a directed cycle. LSTM units and Gated Recurrent Units (GRUs) are variations type of RNN. Thus, Recurrent Neural Networks (RNNs) are employed for processing speech signals [3].

Kurdish is an Indo-Iranian branch of Indo-European languages that are spoken by about 40 million People in Western Asia, primarily in Iraq, Turkey, Iran, Syria, Armenia, and Azerbaijan [3]. Kurdish contains several dialects, as well as its grammatical system and extensive vocabulary [4], [5].

Central Kurdish (also known as Sorani) and Northern Kurdish are the two most widely spoken dialects of Kurdish (also called Kurmanji). Zazaki and Gorani are two further dialects spoken by smaller groups (also known as Hawrami). Kurmanji is the Kurdish language spoken in northern Kurdistan (in Turkey, Syria, and northern Iraq) and written in the Latin (Roman) alphabet; it is also supported by Google Speech Recognition. The Sorani dialect is spoken primarily in the southeast, including Iran and Iraq, and is written in a modified variant of the Arabic alphabet. There is no data for the Sorani dialect in Google Speech Recognition [6].

In [7] mention that, Kurdish is hampered by a lack of resources to support its computational processing needs. Only a few attempts to develop voice recognition resources for the Kurdish language have been made thus far, necessitating the creation of a dataset for their research.

The major contribution of this work is design and implementation of a straightforward hybrid speech to text model for Kurdish (Sorani) that comprises three CNN layers and three (GRUs) layers, this combination in the proposed model architecture produced results that were more accurate.

The rest of this paper is organized as follows: Section two reviews the related works. Section three is the data collections workflow. Section four presents the model architecture and proposed method. In Section five, results are discussed, and finally, the conclusion is in Section 6.

2. LITERATURE REVIEW

Few attempts have been made to recognize Kurdish speech, this review focused on first: those papers in low resources languages (Arabic, Persian and Kurdish), and how audio datasets are built/collected with. Second: CNN and RNN techniques used for recognition is concerned. Kurdish character recognition has received some recent research such as [8]-[10]; however, our work on speech recognition is still in its early stages. The first attempts for Kurdish speech recognition in [7] which presents a dataset extracted from Sorani Kurdish texts from grades one to three of primary school in Iraq's Kurdistan Region. The first attempts for Kurdish speech recognition in [7] which presents a dataset extracted (BD-4SK-ASR) from Sorani Kurdish texts from grades one to three of primary school in Iraq's Kurdistan Region, which contains 200 sentences. Using CMUSphinx to create ASR, narrated by a single speaker using Audacity software at a sampling rate of 16000 and a 16-bit rate mono single channel. After that, another attempts for Kurdish language arise in [11] created a dataset for their work in Kurdistan, Iran, and used Kalditoolkit to develop the identification engine with SGMM and DNN algorithms for the aquatic model. The authors presented WER of Jira ASR system for different topics (SGMM model trained and evaluated by Office data) which are (General: 4.5%, Sport: 10.0%, Economic: 10.9, Conversation: 11.6%, Letter: 11.7%, Politics: 13.8%, Social: 15.3%, Novel: 16.0%, Religious: 16.2%, Scientific/Technology: 17.1%, and Poet: 25.2).

For isolated word recognition, some Arabic papers been reviewed. In [12] proposed, an Arabic digit classification system using 450 Arabic spoken digits. Based on a speaker-independent system, the accuracy was around 93%, the system is based on combining wavelet transform with linear prediction coding LPC method to extract the feature and the probabilistic neural network PNN for classification.

The work by [13] employed Sphinx technologies to recognize solitary Arabic digits with data provided by six different speakers. The system achieved an 86.66%-digit recognition accuracy, examine the use of a distributed word representation and a neural network for Arabic speech recognition. Furthermore, the neural network model allows for robust generalization and improves the ability to combat data sparseness. The inquiry approach also comprises a variety of neural probabilistic model configurations, an n-gram order parameter experiment, output vocabulary, normalization method, model size, and parameters. The experiment was carried out on Arabic news and discussion broadcasts.

Then, in [14] utilized an LSTM neural network for frame-wise phoneme classification on the Farsdat data set, and in [15], they employed a DLSTM with a CTC output layer for Persian phoneme recognition on the same data set.

The rest of this review focused on papers that used CNN, RNN, and GRU or combining of these techniques.

In [16] a significant study was reported. The authors utilized a deep Recurrent Neural Networks (RNN) model that was end-to-end with appropriate regularization. On the TIMIT phoneme recognition benchmark, they found that RNN, namely, Long Short-Term Memory (LSTM), had a test error of 17.7%.

There are, nevertheless, several studies underway to build computational tools for the Kurdish language. In [15] collects a tiny corpus named corpus of contemporary Kurdish newspaper texts (CCKNT), which contains 214K Northern Kurdish dialect terms. Pewan text corpus for Central Kurdish and Northern Kurdish was collected from two online news organizations. The Pewan corpus contains around 18 million tokens for the Central Kurdish dialect and approximately 4 million tokens for the Northern Kurdish dialect. This corpus serves as a validation set for information retrieval applications [17].

In [18], they offered a speech-to-text conversion strategy for the Malayalam language that employs deep learning techniques. For the training, the system is looking at 5–10 solitary words. Mel-frequency cepstral coefficients are acquired for the preprocessing phase. HMM is used to identify the speech and training after the preprocessing, syllabification, and feature extraction procedures. The LSTM was used to construct a speech recognition system based on ANN. The system has a 91% accuracy.

A recurrent neural network approach called LSTM to distinguish individual Bengali words was used in [19] The model is a two-layer deep recurrent neural network with 100 LSTM cells in each layer, 30 unique phonemes are detected, the last layer is a SoftMax output layer with 30 units, and the data set was used with a total of 2000 words. Fifteen different male speakers contributed to the audio speeches. Making a 75:12.5:12.5 split of the dataset for training, validation, and testing purposes. The test run yielded a phoneme detection error rate of 28.7% and a word detection error rate of 13.2%.

In [20] revised standard GRUs for phoneme recognition Purposes and proposed that Li-GRU architecture is a simplified version of a standard GRU, in which the reset gate is removed and ReLU activations are considered, this research worked with (TIMIT, DIRHA, CHiME, TED) corpus Li-GRU outperforms GRU in all the considered noisy environments, with achieving higher performance the bus (BUS) environment (the noisiest) relative improvement of 16% against the relative improvement of 9.5% observed in the street (STR). WER % calculated for DIRHA corpus in real part for (MFCC = 27.8, FBANK = 27.6, fMLLR = 22.8).

The study in [21] employed LSTM and two datasets in this project for Arabic speech recognition. The 1-digit dataset consists of 8800 tokens with a sampling rate of 11025 Hz, and it was created by asking 88 Arabic native speakers to repeat all digits 10 times. 2-TV command dataset: 10000 tokens for 10 TV commands at a sampling rate of 16000 Hz are contained in this dataset; finally, the author reached over 96% accuracy.

The author proposed four different model structures for speech Emotion recognition in [22], which are Model-A (1D CNNs-FCNs), Model-B (1D CNNs-LSTM-FCNs), Model-C (1D CNNs-GRU-FCNs), and Ensemble Model-D, which combines Model-A, Model-B, and Model-C, adding LSTM, and GRU after CNN blocks in models B and C results in increased accuracy (TESS) In total, there are 2800 audio files with 200 target words. 2-RAVDESS audio files have a resolution of 1440 pixels and a sample rate of 48 kHz. 3-SAVEE has a total of 1920 samples. 4-EMO-DB Berlin it contains 535 German-language audio recordings. CREMA-D is the fifth step in the CREMA-D process. It makes use of 7442 records.

3. DATA COLLECTION

This section will discuss how to collect data and which type of data should be collected. Those data were gathered through official administration papers in the University of Sulaimani College of Science, which totaled 500 different words and were collected by 30 speakers, 13 are female and 17 are male.

3.1. Kurdish Sorani Speech (KSS) Dataset

As mentioned before, there is no available dataset in Kurdish Sorani, which lead us to make our dataset for this research work here are the details of workflows, choosing individual words from the governmental worksheet, and arranging them in 30 to 50 words in each paragraph. Four hundred words were read by 30 volunteers and the last 100 words were read by two different male and female readers; the total number of words reached 500 words. There were 30 volunteers in total, with 14 males and 16 females, 9 from family, and 21 from universities. The volunteers' ages ranged from 20 to 40.

The volunteer was asked to read the paragraph as individual words, which means between each word makes silence for at least 0.2 s. Some speakers were asked to read each paragraph 1 time, which leads to 2–3 min' duration of each file, but some others were asked to read each same word 3–5 times the duration of these files reached 5–7 min.

3.2. Recording Circumstance

KSS data sets were collected in two environments office and home with two recording devices that Table 1 – indicates all information needed for data collection.

The dataset consists of 500 words of numeric numbers and formal words. However, from "١١" in Sorani reading "یازده" or "یانزه" in Latin reading "Yazde" or "Yanze" which is eleven in numeric number, to "٢٠" in Sorani reading "بیست" in Latin reading "Bîst" which is twenty numeric number, and also for (١٠٠٠،١٠٠،٩٠،٨٠،٧٠،٦٠،٥٠،٤٠،٣٠ ،١٠٠٠٠٠٠) as the same above for each numeric number, in total 41 tokens, the rest of 461 words containing formal words, weekdays, Kurdish month name, Kurdish pronoun, prefixes, and suffixes, and there are some words in the Kurdish language used to join two same or different words or sentences like (وە ، و ، کە، ی، یان، بۆ، لە), Table 2 shows part of the dataset.

3.3. Recording Technology

This section will discuss the property of the application that is used for recording speech and the recording conditions being explained.

For recording sounds in an office environment, Audacity was utilized since it is a free, easy-to-use, multi-track audio editor and recorder for Windows, MAC OS, GNU/Linux, and other operating systems that can also export sound files as MP3s (MP3, OGG, and WAV).

Using this application, need to set up some recording conditions, that could work properly with the deep learning model, these conditions are described below. using both mono recording channel and stereo recording channel and sample rate of 16000 Hz as it is near to the normal human sound and also using 44100 Hz after converting it to 16000 Hz.

TABLE 1: The information on data collection

Title	Value
Dataset name	Kurdish Sorani Speech
Office Recording Device model	Laptop: - DELL (Latitude E5450) - LENOVO (20ARS0YB08)
No. of Speakers	30
No. of isolated words	501
No. of recorded sound	18,799 sound files (utterance)
Frequency	16000 Hz and 44100 Hz
Recording Channels	Mono and Stereo
Sound Files format	MS Wav (.wav)
Sampling Resolution	32-bit

3.4. Data Preprocessing

After collecting data as mentioned before, the following steps for data preprocessing.

1. Splitting individual words from sound files and saving each one as a new.wav file about the approximately 1-s duration for each one, using a model called "pydub" that can work with audio files, this library can play, split, merge, and edit.wav audio files.
2. After making chunks of the dataset facing many challenges, one of the challenges is appearing some sounds during recording like breathing in loud sounds "uhhh" the program treats as a separate word, should listen to each sound file carefully and discard these as an un-speech sound. Furthermore, some speakers make "umm" or "uh" sounds before or after reading words; in this case, this part has been removed from the entire speech, as a result, it does not affect the dataset.
3. During splitting sound into small chunks (separate words), these small files are read by a package for audio analysis like music generation and ASR. Improves building blocks necessary to create music information retrieval system. Mainly retrieves numerical Numpy array, which represents sound data. Moreover, sampling rate SR is the number of samples taken per second. By default, samples the file at a sampling rate of 22050 Hz, this sample rate could be overridden to any desired SR (8000, 11025, 16000, 22050, 32000, 44100, etc.) Hz. SR = number of samples per second. Taking from a continuous signal to make a discrete or digital signal, choosing a 16000 sample rate.
4. In Fig. 1, easily note that the speech part is ended in 1.2 s and mentioned before in challenge two the 0.2 s, after

TABLE 2: Samples of dataset

Sorani reading style	Latin reading style	Meaning in english
سفر	Sifr	Zero
یەک	Yek	One
دوو	Dû	Two
سێ	Sê	Three
چوار	Çwar	Four
پێنج	Pênç	Five
شەش	Şeş	Six
حەوت	Hewt	Seven
هەشت	Heşt	Eight
نۆ	No	Nine
دە	De	Ten
زانکۆ	Zanko	University
خوێندکار	Xwendkar	Student
یاریدەدەر	Yaridadar	Assistant
پزیشکی	Pzishky	Medicine
ئەنجومەن	Anjuman	Council
زانست	Zanst	Science

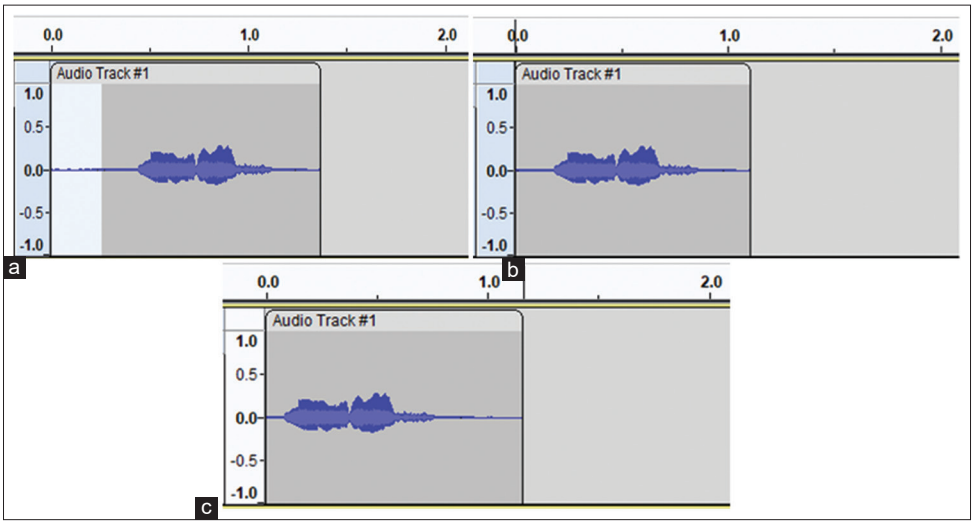


Fig. 1. The speech signal (a) with silence part, in the beginning, (b) with the removed silent part, and (c) with silence part in the end.

1.0 s will be removed and the speech will be unclear. In this case, should remove the silence part from the beginning of the speech, but if the silent part is after the speech, do nothing. The duration of the speech signal is 1.0 s, after that time, it does not matter and fixed it before in challenge two.

5. In some cases, when the word contains two or more sections like "هاوپێچ" which is mean "attach" the reader read it separately, the program treats it as a separate word which makes it mean less "پێچ" "هاو," fixed this problem manually by combining these sections.
6. In opposite to challenge four, in some cases, readers read two different words without any silence between them, and the program selected it as one word, also fixed it manually by separating them like "هەرێمی کوردستان" meaning "Kurdistan Region," as shown in Fig. 2.

1. The format of sound data retrieved from Viber was. m4a which is not supported by Audacity, should convert to a.wav file and also change its sampling rate from 48000 Hz to 16000Hz the size of the file reduced for instance a file size of 7.60 MB becomes 3.8 MB.
2. To prepare the data for fitting into a model, down sampling was applied to the recorded sound files, resulting in 8000 samples per word.

3.5. Data Augmentation (DA)

DA is the method of applying minor modifications to our original training dataset to produce new artificial training samples. There are many types of DA which are time wrapping, frequency masking, time masking, noise reduction, etc.

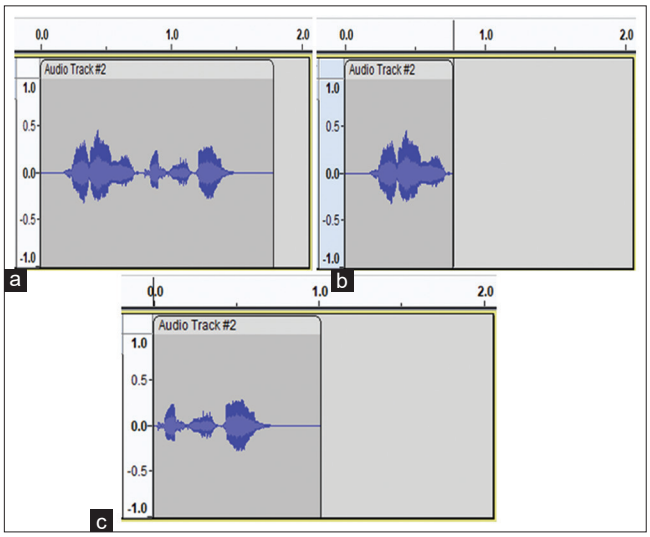


Fig. 2. Separate speech signal to individual words, (a) single chunk with two words, (b) separate first word, and (c) separate second word.

As in [23] used Additive White Gaussian Noise, pitch shifting, and stretching of the signal level. In the proposed work, since the number of speech utterance records in each class is relatively low, this study performs one type of audio DA, which is noise reduction.

4. METHODOLOGY

Both CNN and RNN have widely used in the speech recognition area and approved satisfactory results, for the Kurdish language using these models was challenging. This study proposed four different architecture models,

model generalizations, changing hyper parameters like batch size [24], and optimizers. Related to the study result concluded with CNN lower accuracy, then decided to go with a combination of CNN with RNN (GRU) to seek higher accuracy than the first model architecture.

4.1. CNN Model

The initial model architecture, the CNN model, was used to train and test the KSS dataset. The model comprises four CNN layers, as shown in Fig. 3, each of which is made up of three basic layers, the first of which is conv_layer. This is the first layer, and it is used to extract various features from the input data by performing mathematical operations between the input data and a filter of a specific size (8 × 8, 16 × 16, 32 × 32, 64 × 64) for each CNN layer. The second one A Pooling Layer is usually followed by a Convolutional Layer in most circumstances. This layer's major goal is to lower the size of the convolved feature map to reduce computational expenses. The third one over fitting happens when a model performs so well on training data that it hurts its performance when applied to new data. A dropout layer is used to solve this problem, in which a few neurons are removed from the neural network during the training process, resulting in a smaller model. After passing a dropout of 0.3, 30% of the nodes in the neural network are dropped out at random.

4.2. RNN (GRU)

The vanishing gradient problem affects RNN during back propagation. Gradients are values that are used to update the weights of a neural network. When a gradient reduces as it back propagates through time, this is known as the vanishing gradient problem. When a gradient value falls below a certain threshold, it no longer contributes much to learning. RNNs can forget what they have seen in longer sequences, because these layers do not learn, resulting in short-term memory. As a solution to short-term memory, LSTMs and GRUs were developed. They have inbuilt devices known as gates that can control the flow of data.

GRUs are a recurrent (RNN) gating technique first introduced in [25]. The GRU is similar to an (LSTM) with a forget gate, as shown in Fig. 4, but it has fewer parameters and lacks an output gate. Its performance in polyphonic music modeling, speech signal modeling, and natural language processing was found to be comparable to that of an LSTM. On certain smaller and less frequent datasets, GRUs have been proven to perform better [26].

4.3. Training Models

The ability of the model to adapt to new previously unseen data derived from the same distribution as the one used to generate the model is referred to as generalization. For instance, adding or removing layers to the current model leads to changing the accuracy of the system. Table 3 presents four different architecture models. After getting results from the Table 3, model (3) was chosen as the best model architecture. The detailed layers of the proposed architecture are shown in Fig. 5.

5. RESULTS AND DISCUSSION

In this section, the result of our experiments is presented in two different algorithms, the first is CNN, and the second is CNN+RNN (GRU), as explained in Table 3. As shown in Table 4 indicate that, each batch size got an accuracy. For the CNN model, architecture concluded the best batch size which is Batch size = 16.

To carry out a thorough investigation and achieve a fair comparison between the varied systems, the research uses the 10:90 and 20:80 approaches for testing and training, respectively, as a way of assessing the proposed systems. The results can then be averaged to compute a single estimation. This is particularly important when carrying out experiments with limited data sources, it is important to be clear that the point of this experiment was to ascertain how much data should reserve for testing. The result is presented in Tables 5 and 6 for the 10:90 splitting dataset and Table 7 for the 20:80 splitting dataset.

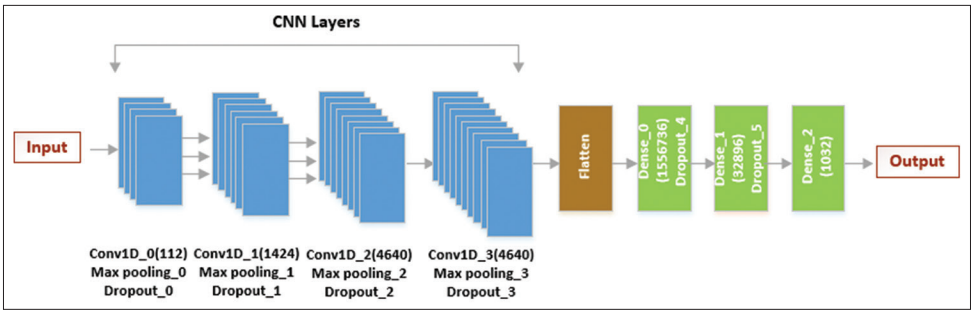


Fig. 3. The CNN architecture model.

After realizing combining three layers of CNN with three layers of RNN which is model, Architecture 3 shows a better result among other 4 architectures, as shown in Table 5, with %90 of the dataset for training the model and %10 for testing the model, after changing hyper parameters like batch size, also using SGD and Adam optimizer, the result shows in Tables 6 and 8.

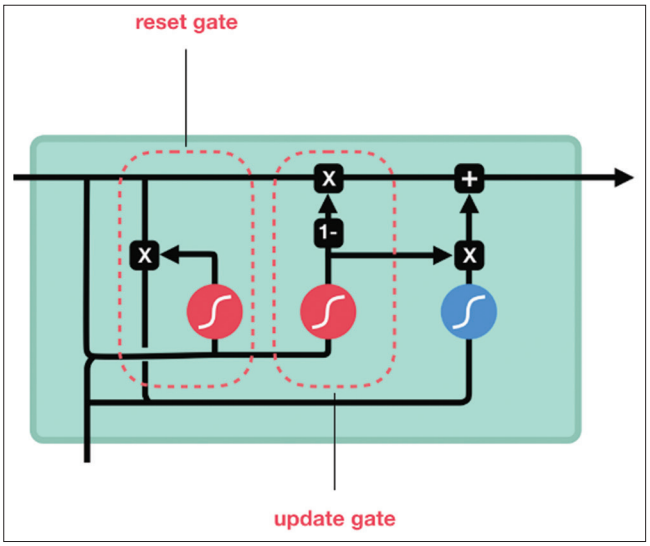


Fig. 4. Gated recurrent unit (Chung *et al*. 2014).

TABLE 3: Different types of model architecture with different layers

Layers	Model 1	Model 2	Model 3	Model 4
Conv_0 ✓	✓	✓	✓	✓
Maxpooling_0	✓	✓	✓	✓
Dropout_0	✓	✓	✓	✓
Conv_1	✓	✓	✓	✓
Maxpooling_1	✓	✓	✓	✓
Dropout_1	✓	✓	✓	✓
Conv_2	✓	✓	✓	×
Maxpooling_2	✓	✓	✓	×
Dropout_2	✓	✓	✓	×
Conv_3	✓	✓	×	×
Maxpooling_3	✓	✓	×	×
Dropout_3	✓	✓	×	×
flatten	✓	✓	×	×
Batch Normalization	×	×	✓	✓
GRU Bidirectional_0,1,2	×	×	✓	✓
Batch Normalization	×	×	✓	✓
Flatten	×	×	✓	✓
Dense_0	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Dense_1	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Dense_2	✓	✓	✓	✓
Dropout	✓	×	×	×

As discussed above indicate that both types of optimizer SGD and Adam could be used for speech recognition as they show a confident result, Adam optimizer reached the result that in batch size (64 and128), shows a better choice as it is accuracy reached (96% and 96%), respectively, which is higher than among batch size (8, 16, and 32) results (92%, 93%, and 72%). On the other hand, the SGD optimizer represents the result in different batch size values which are (16, 32, 64, and 128), but only in (32) reaches the result to (90%). This experiment discovered that the

TABLE 4: CNN model batch size and accuracy number

Batch size	Accuracy %
8	0.51661
16	0.61993
64	0.58672
77	0.47
99	0.40

TABLE 5: The accuracy for each model

Layers	Model 1	Model 2	Model 3	Model 4
Accuracy	0.61	0.88	0.92	0.87

TABLE 6: Using different batch size with ADAM optimizer

Optimizer	Batch size	Epochs	Accuracy %
Adam	8	31	0.92074
Adam	16	59	0.93738
Adam	32	58	0.7278
Adam	64	60	0.9601
Adam	128	69	0.96436

TABLE 7: The effect of splitting dataset to 20:80 on accuracy

Optimizer	Batch size	Epochs	Accuracy %
Adam	8	62	0.89734
Adam	16	59	0.94176
Adam	32	40	0.93697
Adam	64	49	0.94495
Adam	128	50	0.94441

TABLE 8: Using different batch size with SGD optimizer

Optimizer	Batch size	Epochs	Accuracy %
SGD	16	96	0.83989
SGD	32	16	0.90160
SGD	64	29	0.01489
SGD	128	31	0.01755

TABLE 9: Comparison with recent works related to proposed method, dataset, accuracy acheivments

Author	Proposed	Dataset	Acc.
Alkhateeb [12]	Arabic digit classification system using probabilistic neural network PNN.	450 Arabic spoken digits	93%
Arun *et al*. [18]	Malayalam speech recognition, The RNN was used.	5–10 solitary words	91%
Zerari *et al*. [21]	Used RNN for Arabic speech recognition.	Consists of 8800 tokens and TV command dataset: 10000 tokens for 10 TV commands	96%
Proposed System	Using CNN with RNN (GRU) for Kurdish word recognition	KSS Dataset that compose of 18799 sound files for 500 formal Kurdish words	96%

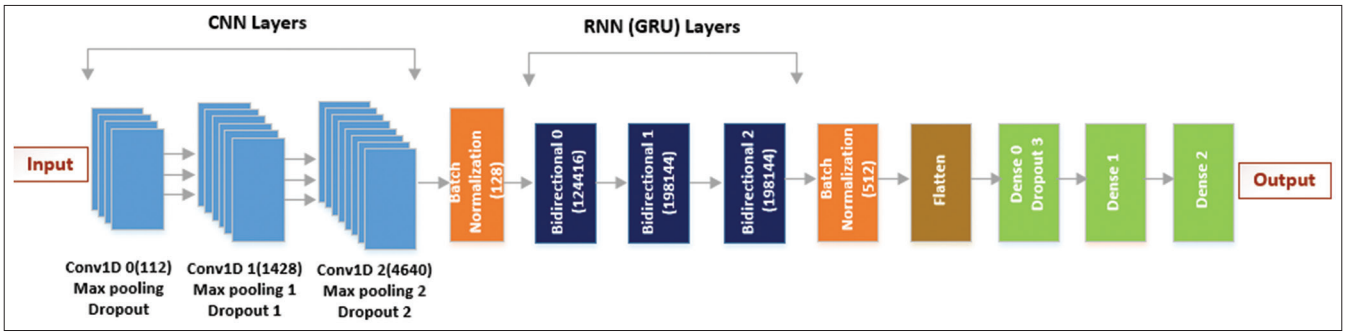


Fig. 5. The proposed architecture CNN_GRU model.

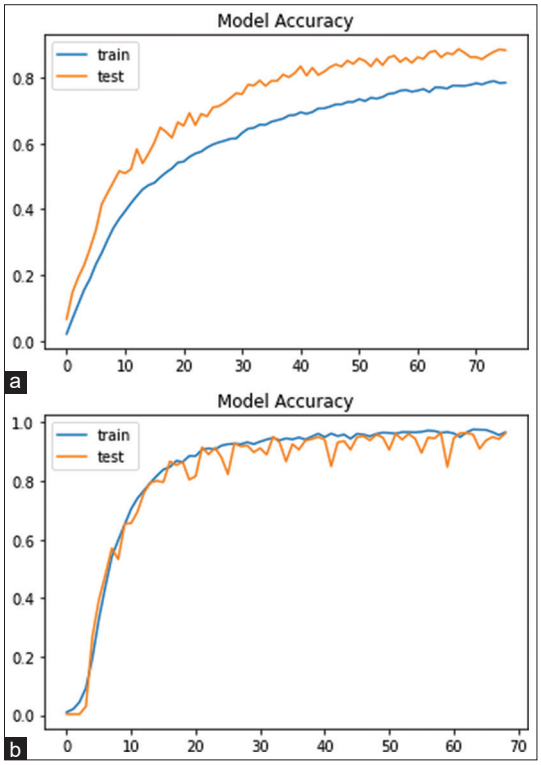


Fig. 6. (a) Accuracy for model CNN. (b) Accuracy for model CNN and RNN (GRU). Horizontal line indicates accuracy, while vertical line indicates number of Epochs.

Adam optimizer is more proper for speech recognition as getting a high accuracy in almost all the tests than the SGD optimizer. Fig. 6 shows the best accuracy for model CNN and RNN (GRU).

Now by changing the splitting dataset to 20:80 for testing: Training with batch size (8, 16, 32, 64, and 128), the results show in Table 7.

By comparison with recent works, the Table 9 indicates the comparison between those papers referenced with proposed method, its dataset and accuracy achieved.

6. CONCLUSION

In this paper, implemented speech recognition for the Kurdish language as well as created a KSS data set for this research purpose. The data set composed of 18799 sound files for 500 formal Kurdish words were read by 30 native Kurdish speakers. The research work designed different model architectures with different parameters using CNN and CNN+ RNN(GRU), the experimental findings indicate that the accuracy of the model increases when three layers of GRU are added to three layers of CNN. The accuracy of the CNN model reaches 61%, but after adding GRU, the accuracy increases dramatically to 96%, providing us with a clear vision for selecting the desired architecture. In the future, intend to improve the quality of Kurdish language materials, as well as utilize state of the art methods such as

metaheuristic optimizer with deep learning, to improve the performance.

REFERENCES

[1] E. Morris. "Automatic Speech Recognition for Low-Resource and Morphologically Complex Languages". *Thesis. Rochester Institute of Technology*, 2021.

[2] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng and J. A. Landay. "Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones". *Journal Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies Archive*, vol. 1, no.4, pp. 1-23, 2017.

[3] M. Assefi, M. Wittie and A, Knight. "Impact of network performance on cloud speech recognition". In: *Proceedings of the 24th International Conference*, pp. 1, 2015.

[4] M. Asseffi, G. Liu, M. P. Wittie and C. Izurieta. "*An Experimental Evaluation of Apple Siri and Google Speech Recognition*". ISCA SEDE Montana State University, Bozeman, 2015.

[5] A. Ganj and F. Shenava. "2-Persian continuous speech recognition software". In: *The First Workshop on Persian Language and Computer*. The 9th Iranian Electrical Engineering Conference, Iran, 2004.

[6] F. A. Ganj, S. A. Seyedsalehi, M. Bijankhan, H. Sameti, S. Zadegan and J. Shenava. "1-Persian continuous speech recognition system". In: *The 9th Iranian Electrical Engineering Conference*, 2000.

[7] A. Qader and H. Hassani. "*Kurdish (Sorani) Speech to Text: Presenting an Experimental Dataset*". arXiv: 1911.13087v1, 2019.

[8] R. Yaseen and H. Hassani. "Kurdish Optical Character Recognition". *UKH Journal of Science and Engineering*, vol. 2, pp. 18-27, 2018.

[9] R. D. Zarro and M. A. Anwer. "Recognition-based online Kurdish character recognition using hidden Markov model and harmony search Eng." *Engineering Science and Technology an International Journal*, vol. 20, no. 2, pp. 783-794, 2017.

[10] A. T. Tofiq and J. A. Hussain. "Kurdish Text Segmentation using projection-based approaches". *UHD Journal of Science and Technology*, vol. 5, no. 1, pp. 56-65, 2021.

[11] H. Veisi, H. Hosseini, M. Amini, W. Fathy and A. Mahmudi. "*Jira: A Kurdish Speech Recognition System Designing and Building Speech Corpus and Pronunciation Lexicon*". ArXiv abs/2102. 07412, 2021.

[12] A. Alkhateeb. "Wavelet LPC with neural network for spoken arabic digits recognition system". *Jordan Journal of Applied Science*, vol. 4, pp. 1248-1255, 2014.

[13] N. Turab, K. Khatatneh and A. Odeh. "A novel arabic speech recognition method using neural networks and gaussian filtering". *IJEECS International Journal of Electrical, Electronics and Computer Systems*, vol. 19, pp. 1-5, 2014.

[14] S. Malekzadeh, M. H. Gholizadeh and S. N. Razavi. "*Persian Phonemes Recognition Using PPNet*". arXiv preprint arXiv: 1812.08600, 2018.

[15] H. Veisi and A. Haji Mani. "Persian speech recognition using long short-term memory". In: *The 21st National Conference of the Computer Society of Iran*. University of Tehran, Iran, 2015.

[16] A. Graves, A. R. Mohamed and G. Hinton. "Speech recognition with deep recurrent neural networks". In: *ICASSP Conference*. Institute of Electrical and Electronics Engineers, Piscataway, 2013.

[17] A. R. Mohamed, G. Dahl and G. Hinton. "Deep belief networks for phone recognition". In: *Nips Workshop on Deep Learning for Speech Recognition and Related Applications*. . IJCA Proceedings on National Conference, USA, 2009.

[18] H. P. Arun, J. Kunjumon, R. Sambhunath and A. S. Ansalem. "Malayalam speech to text conversion using deep learning". *IOSR Journal of Engineering* (*IOSRJEN*), vol. 11, no. 7, pp. 24-30, 2021.

[19] M. M. H. Nahid, B. Purkaystha and M. S. Islam. "Bengali speech recognition: A double layered LSTM-RNN approach". In: *Procceding 20th Institute of Communication Culture Information and Technology*, pp. 1-6, 2017.

[20] M. Ravanelli, P. H. Brakel, M. Omologo and Y. Bengio. "Light gated recurrent units for speech recognition". *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, pp. 92-102, 2018.

[21] N. Zerari, S. Abdelhamid, H. Bouzgou and C. Raymond. "Bidirectional deep architecture for Arabic speech recognition". *Open Computer Science*, vol. 9, pp. 92-102, 2019.

[22] R. Ahmed, S. Islam, A. K. M. Muzahidul Islam and S. Shatabda1. "An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition". arXiv: 2112.05666, 2021.

[23] C. Huang, G. Chen, H. Yu, Y. Bao and L. Zhao. "Speech emotion recognition under white noise". *Archives of Acoustics*, vol. 38. pp. 457-463, 2013

[24] I. Kandel and M. Castelli. "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset". *ICT Express*, vol. 6, no. 4, pp. 312-315, 2020.

[25] K. Cho, B. V. Merrienboer, D. Bahdanau and Y. Bengio. "*On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*". arXiv: 1409.1259v2, 2014.

[26] J. Chung, C. Gulcehre, K. Cho and Y. Bengio. "*Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*". arXiv: 1412.3555v1, 2014.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

COVID-19 Disease Detection Based on Machine Learning and Chest X-Ray Images

Ramyar A. Teimoor, Mihran A. Muhammed

Department of Computer, College of Science, University of Sulaimani, Iraq

ABSTRACT

Due to increasing population, automated illness identification has become a critical problem in medical research. An automated illness detection framework aids physicians in disease diagnosis by providing precise, consistent, and quick findings, as well as lowering the mortality rate. Coronavirus (COVID-19) has expanded worldwide and is now one of the most severe and acute disorders. To avoid COVID-19 from spreading, making an automatic detection system based on X-ray chest pictures ought to be the quickest diagnostic alternative. The goal of this research is to come up with the best model for detecting COVID-19 diagnosis with the greatest accuracy. Therefore, four models, Convolutional Neural Networks, Residual Network 50, Visual Geometry Group 16 (VGG16), and VGG19, have been evaluated using the same images preprocessing method. In this study, performance metrics include accuracy, precision, recall, and F1 scores are used for evaluating proposed method. According to our findings, the VGG16 model is a viable candidate for detecting COVID-19 instances, because it has highest accuracy; in result overall accuracy of 98.44% in training phase, 98.05% invalidation phase and 96.05% in testing phase is obtained. The results of other performance measurements are shown in the result section, demonstrating that the majority of the approaches are more than 90% accurate. Based on these results, radiologists may find the proposed VGG16 model to be an intriguing and a helpful tool for detecting and diagnosing COVID-19 patients quickly.

Index Terms: COVID-19, Convolutional Neural Networks, Residual Network 50, Visual Geometry Group 19, Visual Geometry Group 16, X-ray Image, Machine Learning

1. INTRODUCTION

Humans have been victims of many pandemics throughout history. The globe is presently dealing with another epidemic and an unseen foe, the new COVID-19 coronavirus. The coronavirus illness (COVID-19) is a global epidemic that has spread rapidly. The first incidence was reported in Wuhan, China, in November 2019. Following reports of a large number of infections in several nations, the "World Health Organization (WHO)" declared the pandemic a public health emergency on March 11, 2020 [1]. According to the WHO, signs of a respiratory infection caused by the virus include inflammation of the lungs, fever, and coughing. Another aggravating issue is that the virus has a high person-to-person transmission ratio [2], The best way to stop and postpone transmission is to keep your distance from people in social situations [3]. More than 150 million individuals had contracted the coronavirus by March 10, 2021, and 3 million people had died as a result of it [1].

Because of its high transmissibility, early identification of the Coronavirus is critical for managing COVID-19. According to Chinese government standards, Reverse Transcription-Polymerase Chain Reactions should primarily be used to identify the presence of the corona virus in respiratory or blood samples using gene sequencing (RT-PCR). The RT-PCR procedure takes 4–6 h to complete

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp126-134

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Ramyar A. Teimoor, Department of Computer Science, College of Science, University of Sulaimanyah, Sulaimanyah, Kurdistan Region, Iraq. ramyar.teimoor@univsul.edu.iq

Received: 01-10-2022 Accepted: 05-11-2022 Published: 21-11-2022

which is a considerable period comparing to how rapidly COVID-19 spreads [4]. In addition, RT-PCR test kits are not only useless, but also hard to come by. As a consequence, a lot of infected patients go unnoticed for a while and unintentionally spread the disease to others. The prevalence of COVID-19 illness will reduce if the condition is detected early enough.

In order to address the ineffectiveness and unavailability of the current COVID-19 testing, other test methodologies have been explored to identify COVID-19 infections. Additionally, radiological imaging computed tomography (CT) methods like computed tomography and X-rays may be used. According to the study [4], a method based on chest CT scans might be a useful tool for locating and counting COVID-19 instances. X-ray pictures have been used by several studies to illustrate different methods for detecting COVID-19. It is crucial to understand that utilizing technology ethically will improve human lives [5]. Therefore, recent advances in technology such as computer vision, machine learning, and deep learning have allowed for the automated detection of a number of diseases in the human body, ensuring intelligent treatment. Deep learning is utilized as a one of the best techniques for training a model for classification and detection COVID-19 [4].

As a result, one of the main objectives of this study is to offer a machine learning-based system for COVID-19 illness identification using 4 distinct deep learning models.

The remaining sections of this paper are included as follows: In section 2, we go through the existing literature on utilizing Machine learning techniques to analyze COVID-19 CXR images. The proposed Model is described in depth in section 3, including the required parameters, pre-trained backbones, and procedural stages. In section 4, categorization results are contrasted in terms of recall, precision, and overall accuracy to evaluate the model's efficacy. Section 5 comes to a close.

2. RELATED WORK

2.1. Ahammed *et al.* [6]

The goal of this research is to see if using machine learning and deep learning techniques on chest X-ray pictures may help detect coronavirus cases. The two chest X-ray datasets were collected from Kaggle and Github and pre-processed using random sampling to create a single dataset. They used a combination of machine learning and deep learning techniques, including convolutional neural networks (CNN) and traditional machine learning. They looked at specificity, fallout rate, and accuracy to see whether they could better identify non-COVID-19 people. Their suggested that CNN model had the best accuracy (94.03%), AUC (95.52%), f-measure (94.03%), sensitivity (94.03%), and specificity (97.01%), as well as the lowest fall out (4.48%) and miss rate (2.98%).

2.2. Saiz and Barandiaran [7]

They provide a rapid technique for detecting COVID-19 in chest X-ray pictures using deep learning techniques. A publicly available dataset of 1500 images of healthy individuals and those with COVID-19 and pneumonia infection is used to train, test, and recommend object identification architecture. The main goal of their strategy is to categorize a patient's COVID-19 case as either positive or negative. Utilizing the SDD300 model, they successfully used deep learning models to identify COVID-19 with a sensitivity of 94.92% and a specificity of 92.00% in their studies.

2.3. Ismael and Şengür [8]

In this study, to categorize COVID-19 and normal chest X-rays, deep learning-based techniques were utilized, including deep feature extraction, fine-tuning of CNN (Visual Geometry Group [VGG]16 and VGG19, Residual Network [ResNet]18, ResNet50 and ResNet10), and end-to-end training of a proposed CNN model. Support vector machine (SVM) is used to classify deep features; linear, quadratic, cubic, and Gaussian kernel functions were employed. The fine-tuning process also utilized the previously stated pre-trained deep CNN models. In this research, an end-to-end training paradigm is presented to suggest a new CNN. There were 180 COVID-19 and 200 non-COVID-19 images in this dataset. The study's performance was measured in terms of classification accuracy. Various local texture descriptors and SVM outcomes were also employed and compared versus alternative deep methods' performance; the deep methods outperformed local texture descriptors for using chest X-rays in COVID-19 classification. The highest accuracy score achieved among this work's findings obtained from the ResNet50 model and SVM classifier using the linear kernel function.

2.4. Zebin and Rezvy [9]

Using a transfer learning process, they classified COVID-19 X-ray chest pictures from two datasets. The classifier

successfully separates those with and without infection from those with COVID-19 and pneumonitis inflammation in the lungs.They used a number of pre-trained CNN, including VGG16, ResNet50, and EfficientNetB0, for this purpose. About 90%, 94.3%, and 96.8% overall detection accuracy were attained for each, respectively. They also developed and improved the COVID-19 class, a minority in their approach, with the use of a generative adversarial architecture (a CycleGAN). To emphasize the parts of the input image that are crucial for predictions for visual interpretation and explanation, they adopted a gradient class activation mapping technique. They suggest that these visualizations might be used to track the areas of the lung that are damaged as the illness expands and becomes worse.

2.5. Gaur *et al*. [10]

Three pre-trained CNN models are examined in this research (EfcientNetB0, VGG16, and InceptionV3) using transfer learning. These particular models were chosen based on their balance of precision and effectiveness with fewer parameters, making them perfect for mobile applications. The dataset for the research was gathered from a number of freely available sources. Performance metrics and deep learning methods are used in this research (accuracy, recall, specificity, precision, and F1 scores). The results showed that the suggested method provided a high-quality model with a sensitivity of 94.79% for COVID-19 and an overall accuracy of 92.93%. According to the study, computer vision design might be employed to provide efficient ways for detection and screening.

2.6. Sitaula and Hossain [11]

This study introduces a brand-new attention-based deep learning model that makes use of the attention module with the VGG-16 to record the relationship in space between the ROIs in CXR pictures. Meanwhile, they build a VGG-16 model that includes the attention module in addition to a suitable convolution layer (fourth pooling layer). This model's accuracy in the VGG-16 suggested model is 79.58%.

3. METHOD

Through the use of machine learning, this work seeks to develop a prediction model for COVID-19 positive or negative, such as CNN, ResNet50, VGG19, and VGG16. The system components' diagram is presented in Fig. 1, each stage will be covered in detail in the following sections.

3.1. Data Collection

At the first step of the proposed system, we need a dataset. The dataset used for model training, data analyzing, augmentation, and description of data. Images of the chest obtained from Kaggle's COVID collection include X-rays of the chest [12], [13].

The original dataset conation images of four different categories which is (COVID, lung opacity, normal, and viral pneumonia). In this study, only (COVID-19 and Normal) categories have been used because the main focus is to determining that the patient has COVID-19 or not.

There are 13,808 pictures in the dataset related to these two categories, (Fig. 2) each category has the following number of images:

- Total COVID images = 3616
- Total Normal images = 10192.

3.2. Data Preprocessing

The next step after data collection is pre-processing, so is carried out using a variety of Python jupyter notebook built-in parameters and functions, the details of pre-processing of each image have been shown in the Table 1.

Figs. 3 and 4 demonstrate the difference before and after image pre-processing.

3.3. Data Splitting

After preprocessing, the dataset has been divided into train, validate and test categories, because each one of them has a different role in the proposed system. After several times of model training and testing, the division range for train, validate, and test sets that provide the best results are shown in Table 2.

3.4. Models Architecture

In this study, 13,808 chest X-ray pictures were utilized to detect COVID-19 infection using VGG-16, VGG-19, ResNet50, and CNN models. The next part provides a short overview of model architecture, followed by an explanation of the proposed models and their specifications.

3.4.1. CNN

The design of CNN aims to resemble the human visual brain. The convolution layer, the pooling layer, and the fully linked layer are the three primary layers that comprise CNN. In CNN Model learning is done by the convolution and pooling layers, while classification is done by the fully connected layers, (Fig. 5) [14].

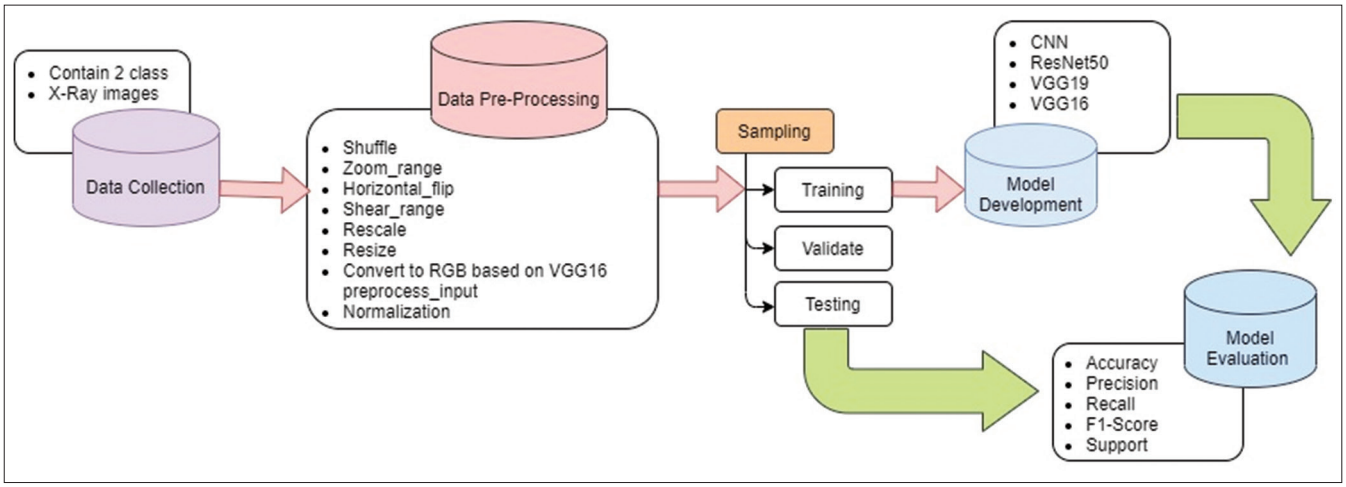


Fig. 1. General diagram of the proposed COVID-19 detection system.

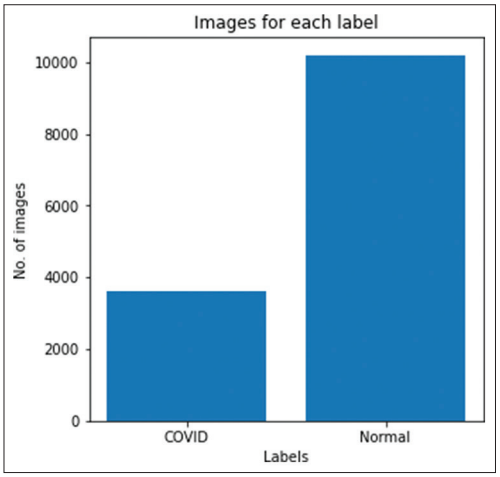


Fig. 2. Number of images in each category.

TABLE 1: Parameter value for pre-processing section

Pre-processing parameter	Value	
Zoom_range_ parameter	0.2	Zooms the image to 20% of its original size
Horizontal_flip_ parameter	True	Flips image horizontally
Shear_range_ parameter	0.2	Shear angle in counter clockwise
Rescale_ parameter	1/255	Rescale image values to 0–1
Resize_ parameter	244×244	Resize image shape
Convert Image to RGB	VGG16_preprocessing VGG19_preprocessing ResNet50_preprocessing	

ResNet50: Residual network 50, VGG: Visual geometry group

3.4.2. ResNet

ResNet stands for residual network in short. Residual learning is the novel phrase that this network presents; as its name suggests. A ResNet variant called ResNet50 has 48 Convolutional layers, one MaxPool layer, and one Average Pool layer. It can do 3.8 × 109 floating-point computations overall [9].

The "University of Oxford's" Simonyan and Zisserman provide the VGG16 Model as a convolutional neural network model in their article; "Very Deep Convolutional Networks for Large-Scale Image Recognition." The model achieves top-5 test accuracy of 92.7% in ImageNet, a dataset with over 14 million images separated into 1000 classes [15]. The model presented to the 2014 ILSVRC was well-known. By gradually substituting several 3 × 3 kernel-size filters for a large number of larger kernel-size filters 11 and 5, respectively, in the first and second convolutional layers, it beats AlexNet.

The input to the cov1 layer of the VGG16 architecture is a 224 × 224 RGB picture. The picture is processed by a series of convolutional layers that employ filters with extremely small receptive 3*3 fields (This is the least size required to correctly capture the left and right, up and down, and center concepts). In additiom, one of the options employs 1 × 1 convolution filters, which linearly change the input channels followed by non-linearity. Convolution stride and spatial padding of the convolution layer input are both set to 1 pixel for 3 × 3 convolution layers to preserve the spatial resolution after convolution. The spatial pooling process uses five max

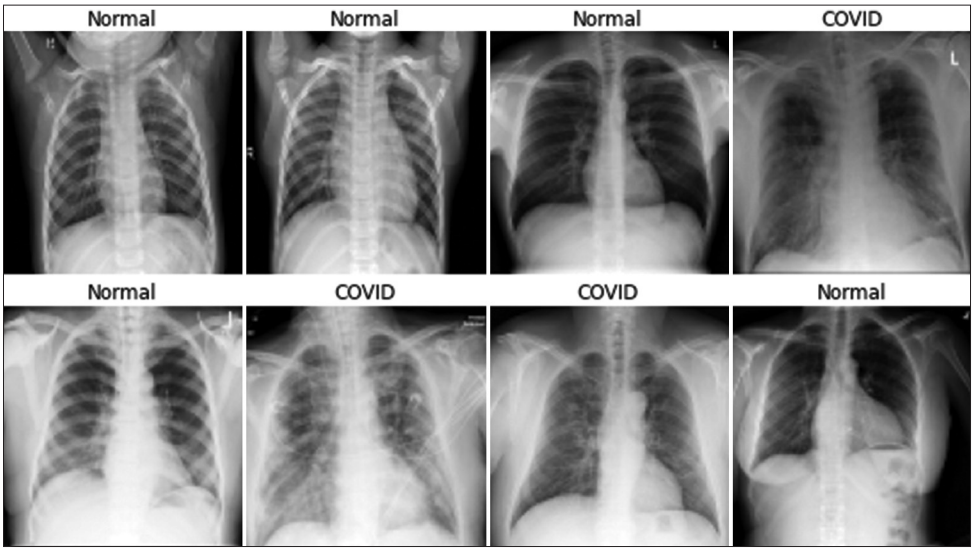


Fig. 3. Images of chest X-rays from the dataset, together with their labels.

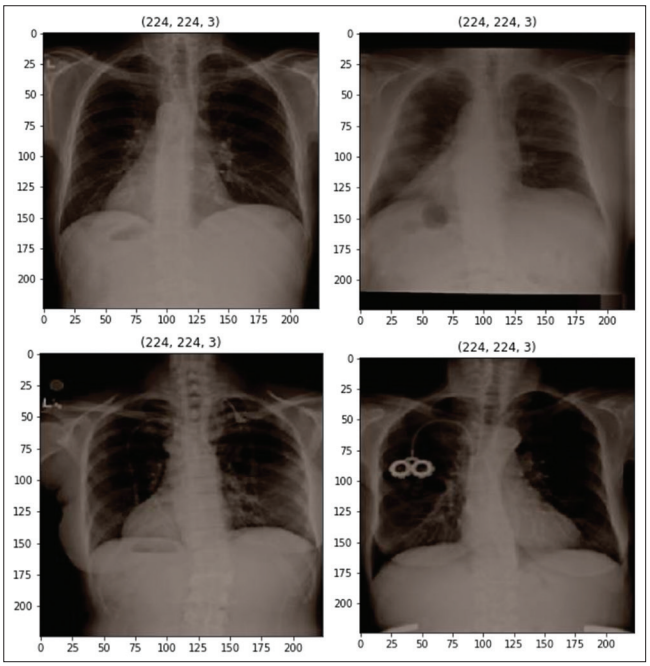


Fig. 4. Images after pre-processing.

pooling layers, which are applied after part of the conv layers. Max-pooling is not always used after convolutional layers. For max-pooling in stride 2, a 2 × 2 is used [15].

In various configurations, a stack of convolutional layers with varied depth is increased with three fully connected layers, and then the following is carried out: Soft-max layer is the bottom layer. All networks construct the fully linked levels

TABLE 2: Number of samples in train, validate, and test set

Sample type	Train	Validate	Test	Total
COVID-19	2416	600	600	3616
Normal	7092	1500	1600	10,192
Total	9508	2100	2200	13,808

in the same way [15]. As a consequence, VGG-16 features three fully connected layers in addition to 13 convolutional layers (Fig. 6) [14].

3.4.3. VGG-19

The VGG-19, on the other hand, consists of three completely connected layers and 16 convolutional layers. Consequently, VGG-19 is seen as a more sophisticated CNN architecture than VGG-16 [14].

3.5. Training Phase

The proposed models were trained based on optimal hyperparameters, one of them is using the Adaptive Moment Estimation (Adam) optimizer [16]. This is the most popular and successful gradient descent optimization technique [17].

Categorical cross-entropy loss function in another hyperparameters that used in this study, it shows that we have multi class to identify so, it is required when we have multi classifications task. Each class's loss is calculated separately [18].

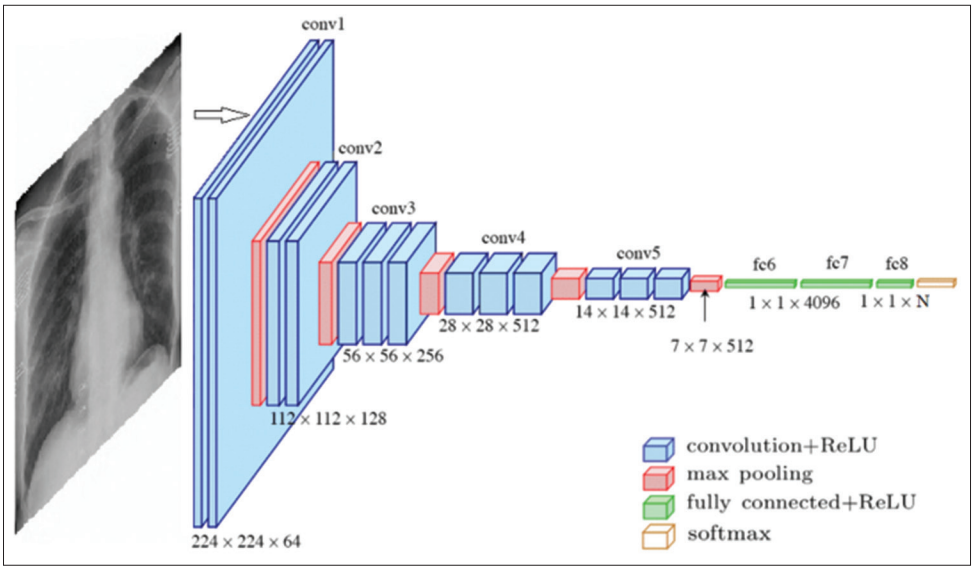


Fig. 5. Convolutional neural networks architecture.

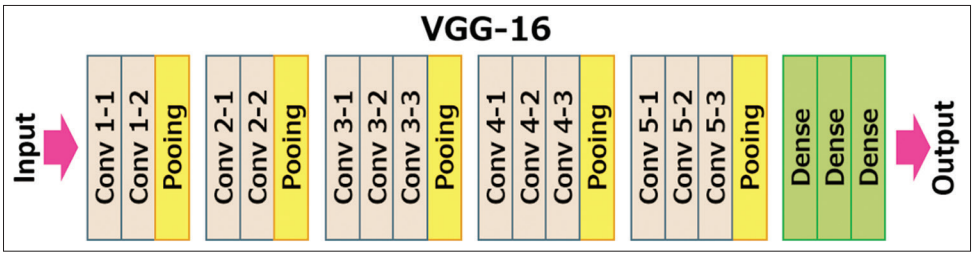


Fig. 6. Visual geometry group 16 model [10].

Table 3 shows other optimal hyperparameters for the CNN, ResNet50, VGG19, and VGG16 architectures that provide the best results.

The model was run on Google Colab since it provides us with free GPU, which is extremely useful for us because, as we all know, training takes a long time, so one of the most important aspects of any study is obtaining results as quickly as possible.

4. RESULTS

The best experiment results are shown in this part to provide how and compare the performance of the recommended VGG16 architecture to the other three models in classifying X-ray pictures into the normal or COVID-19 classes. A total of 13,808 images were used by both classes. Approximately 69% of the photos were utilized for training, and the remaining 31% were divided nearly equally between testing and validating.

TABLE 3: CNN, ResNet50, VGG19, and VGG16 Model specifications

Model argument	Value
Activation function	Sigmoid
Monitor	Val_accuracy
Verbose	1
Epoch	100
Loss function	Categorical cross entropy
Optimizer	Adam
Metrics	Accuracy
Steps_per_epoch	10
Validation_steps	32

CNN: Convolutional neural networks, ResNet50: Residual network 50, VGG: Visual geometry group

The overall accuracy of all four models, namely, (CNN, ResNet50, VGG19, and VGG16) is presented in Table 4. According to these results, all models have an overall accuracy better than 95%, except for the CNN which has lowest accuracy of 80.04%; while VGG16 gives the highest accuracy 98.44, 98.05, and 96.05 for train, validation, and

test, respectively. Thus, in comparison to other works that are referenced in related works, our model has higher accuracy.

Each model is evaluated in addition to accuracy using performance indicators such as precision, recall, and F1-score. Table 5 summarizes the outcome of these metrics for all models.

It can be seen from the numbers in this table, VGG16 earned the highest for practically all metrics, with the exception of the precision of normal class, which is better with ResNet50. As a result, VGG16 could be used as a better-balanced classification model for COVID-19.

A confusion matrix is another way to assess relevance of the models for this problem. Fig. 7 illustrates confusion matrix for all models. A total of 306 of the 2200 test set images are misclassified by CNN. Whereas COVID-19 was detected in 258 normal cases, and 48 COVID-19 as normal case. Resnet50, VGG19 misclassify 125 and 146 images, respectively, for both categories, when VGG16 misclassifies only 81 images.

TABLE 4: Accuracy result for all models

Model	Accuracy%		
	Train	Validate	Test
CNN	83.44	83.3	80.04
ResNet50	97.5	96.43	93.68
VGG19	95.63	94.23	89.95
VGG16	98.44	98.05	96.05

CNN: Convolutional neural networks, ResNet50: Residual network 50, VGG: Visual geometry group

TABLE 5: Precision, recall, and F1-score result for all models

Models	Sample Type	Precision	Recall	F1-score
CNN	COVID	0.88	0.57	0.69
	Normal	0.86	0.97	0.91
	Macro_avg	0.87	0.77	0.80
	Weighted_avg	0.86	0.86	0.85
ResNet50	COVID	0.86	0.95	0.90
	Normal	0.98	0.94	0.96
	Macro_avg	0.92	0.94	0.93
	Weighted_avg	0.95	0.94	0.94
VGG19	COVID	0.87	0.88	0.88
	Normal	0.96	0.95	0.95
	Macro_avg	0.92	0.92	0.92
	Weighted_avg	0.93	0.93	0.93
VGG16	COVID	0.97	0.90	0.93
	Normal	0.96	0.99	0.98
	Macro_avg	0.96	0.94	0.95
	Weighted_avg	0.96	0.96	0.96

CNN: Convolutional neural networks, ResNet50: Residual network 50, VGG: Visual geometry group

Thus, the confusion matrix values are likewise consistent with earlier measures. CNN gives the highest rate of misclassification. On the other hand, while ResNet50 and VGG19 provide promising results for COVID-19 detection, VGG16 provides better accurate outcome in this regard.

For both test and validation sets, accuracy and loss of the VGG16 architecture are displayed in Fig. 8 for accuracy and Fig. 9 for loss. We can see that epoch 64 had the highest accuracy and lowest loss.

Fig. 10 shows two randomly selected images from each class with their counterpart heat maps, which correctly classified by the VGG16 model.

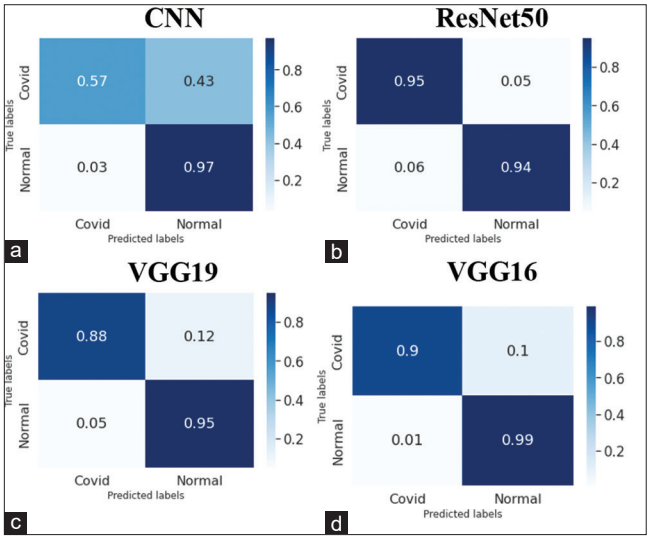


Fig. 7. (a-d) Confusion matrix for (CNN, ResNet50, VGG19, and VGG16). CNN: Convolutional neural networks, ResNet50: Residual network 50, VGG: Visual geometry group.

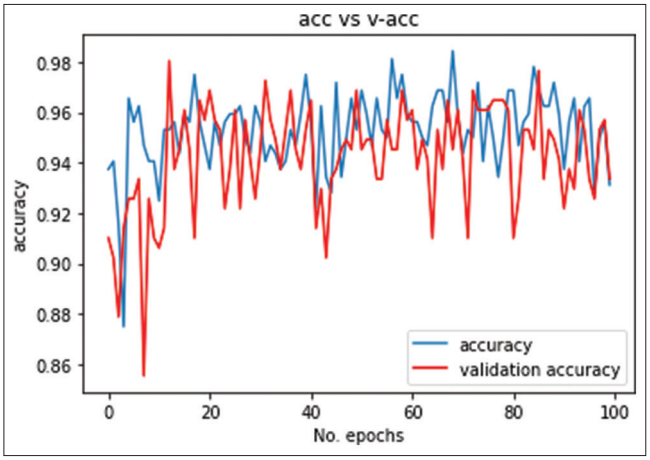


Fig. 8. Visual geometry group 16 model accuracy and validation chart.

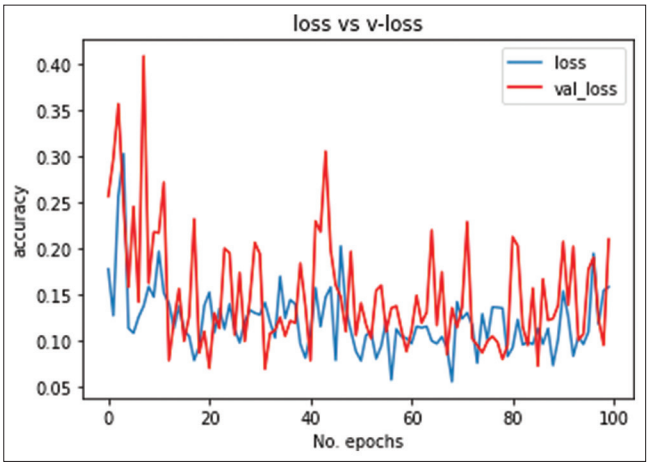


Fig. 9. Visual geometry group 16 model loss and validation loss chart.

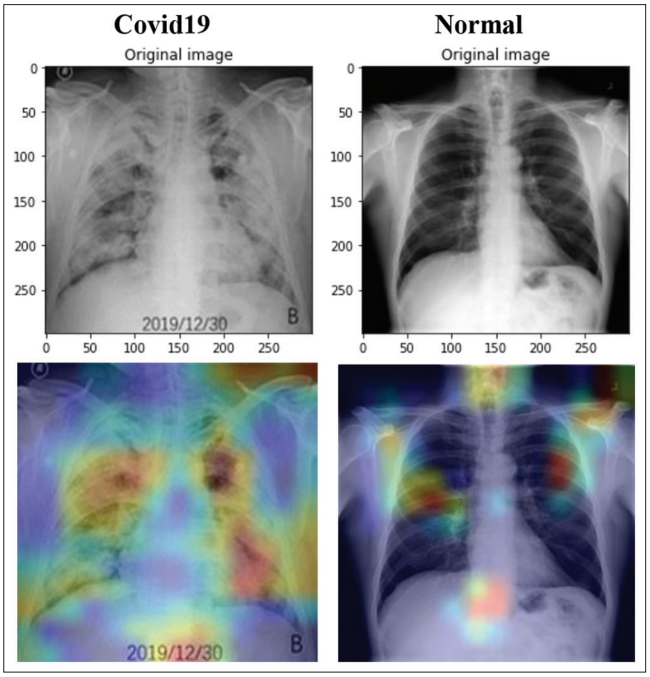


Fig. 10. Heat map images after detection by visual geometry group 16.

5. CONCLUSION

COVID-19 disease is affecting the whole world, the number of infections increasing every day, unfortunately the death cases as well. This work aimed to demonstrate the use of machine learning to a chest X-ray picture to assist the health department in quickly identifying all COVID-19-positive cases and halt the spread of the disease. As the result four alternative models, CNN, ResNet50, VGG16, and VGG19 were used and tested with various parameters to get the best outcome. The results indicate that VGG16, which contains 13 layers of convolutional layers and three fully connected layers, was the best model in our work. Since it has the highest accuracy compared to other models, it also aids physicians in accurately diagnosing the disease. As it is known that COVID-19 disease manifests itself at various phases and with various patterns; this characteristic could be addressed in the future researches. Furthermore, it will be easier for radiologists to have a graphical user interface application to do the real-world tests in hospitals.

REFERENCES

[1] A. Rehman, T. Saba, U. Tariq and N. Ayesha. "Deep learning-based COVID-19 detection using CT and X-ray images: Current analytics and comparisons". *IT Professional*, vol. 23, no. 3, pp. 63-68, 2021.

[2] M. Maia, J. S. Pimentel, I. S. Pereira, J. Gondim, M. E. Barreto and A. Ara. "Convolutional support vector models: Prediction of coronavirus disease using chest x-rays". *Information*, vol. 11, no. 12, pp. 1-19, 2020.

[3] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar and Y. V. R. Pawan. "Analysis, prediction and evaluation of COVID-19 datasets". *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 2199-2204, 2020.

[4] M. Z. Islam, M. M. Islam and A. Asraf. "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images". *Informatics in Medicine Unlocked*, vol. 20, p. 100412, 2020.

[5] R. A. Teimoor and A. M. Darwesh. "Node detection and tracking in smart cities based on internet of things and machine learning". *UHD Journal of Science and Technology*, vol. 3, no. 1, pp. 30-38, 2019.

[6] K. Ahammed, M. S. Satu, M. Z. Abedin, M. A. Rahaman and S. M. S. Islam. "Early detection of coronavirus cases using chest X-ray images employing machine learning and deep learning approaches". *medRxiv*, Vol. 2, p. 2020.06.07.20124594, 2020.

[7] F. Saiz and I. Barandiaran. "COVID-19 detection in chest X-ray images using a deep learning approach". *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, p. 4, 2020.

[8] A. M. Ismael and A. Şengür. "Deep learning approaches for COVID-19 detection based on chest X-ray images". *Expert Systems with Applications*, vol. 164, p. 114054, 2021.

[9] T. Zebin and S. Rezvy. "COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization". *Applied Intelligence*, vol. 51, no. 2, pp. 1010-1021, 2021.

[10] L. Gaur, U. Bhatia, N. Z. Jhanjhi, G. Muhammad and M. Masud. Medical image-based detection of COVID-19 using deep convolution neural networks. *Multimedia Systems*, vol. 27, pp. 1-10, 2021.

[11] C. Sitaula and M. B. Hossain. "Attention-based VGG-16 model for COVID-19 chest X-ray image classification". vol. 19, pp. 2850-2863, 2021.

[12] "COVID-19 Radiography Database." Available from: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[13] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.

A. Kadir, Z. B. Mahbub, *et al*., "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665-132676, 2020.

[14] R. Mohammadi, M. Salehi, H. Ghaffari and A. A. Rohani. "JBPE_ Volume 10_Issue 5_Pages 559-568.pdf". vol. 2019, pp. 559-568, 2020.

[15] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1-14, 2015.

[16] H. Swapnarekha, H. S. Behera, D. Roy, S. Das and J. Nayak. "Competitive deep learning methods for COVID-19 detection using X-ray images". *Journal of The Institution of Engineers* (*India*)*: Series B*, 2021.

[17] C. Ouchicha, O. Ammor and M. Meknassi. "CVDNet: A novel deep learning architecture for detection of coronavirus (COVID-19) from chest x-ray images". *Chaos, Solitons and Fractals*, vol. 140, 2020.

[18] J. E. Luján-García, M. A. Moreno-Ibarra, Y. Villuendas-Rey and C. Yáñez-Márquez. "Fast COVID-19 and pneumonia classification using chest X-ray images". *Mathematics*, vol. 8, no. 9, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Enhanced Single Image Dehazing Technique based on HSV Color Space

Mohammad Khalid Othman[1], Alan Anwer Abdulla[2,3]

[1]Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq, [2]Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq, [3]Department of Information Technology, University College of Goizha, Sulaimani, Iraq

ABSTRACT

The clarity of images degrades significantly due to the impact of weather conditions such as fog and haze. Persistent particles scatter light, attenuating reflected light from the scene, and the dispersed atmospheric light will mix with the light received by the camera affecting image contrast in both outdoor and indoor images. Conventionally, the atmospheric scattering model (ATSM) is a model often used to recover hazy images. In ATSM, two unknown factors/parameters must be estimated: Airlight and scene transmission. The accuracy of these estimations has a significant influence on the dehazed image quality. This paper focuses on the first parameter. It introduces a new technique for estimating the airlight based on the HSV color space. The HSV color space is utilized to identify the haziest opaque area in the image. Consequently, the amount of airlight in the selected area is calculated. To assess the effectiveness of the suggested approach, the well-known dataset, RESIDE SOTS, has been used that contains two parts; namely, SOTS-indoor and SOTS-outdoor. Each of dataset includes 500 images. Experimental findings show that the suggested approach outperforms the existing techniques in terms of peak signal-to-noise-ratio and structural similarity index`.

Index Terms: Dehazing, Atmospheric Scattering Model, Color Spaces, HSV, Haze

1. INTRODUCTION

Digital image processing (DIP) is important in many fields including medical image processing, image in-painting, pattern recognition, biometrics, content-based image retrieval, image dehazing, and multimedia security [1], [2]. In DIP, visibility is a major problem that image vision-based systems must deal with. Weather condition makes a scene less visible, which has an impact on how well outdoor image processing-based systems work such as detection and recognition of objects, visual surveillance, traffic monitoring, intelligent transportation, and etc. [3]. The camera captures a small portion of the light reflected directly from the surface of an object as well as a significant portion of the light reflected by the atmosphere. A light that is reflected from the surface of an object is scattered and absorbed by atmospheric particles [4]. In bad weather, these scattering and absorption increase causing the irradiance to be measured incorrectly [5]. In addition, inclement weather makes the images and videos degrade and this leads objects lose their contrast and visibility [6]–[9]. Applications based on computer vision operate perfectly when the input is noiseless. Many applications perform unsuccessfully in bad weather due to the fading images and videos. Therefore, image dehazing is required for computer vision-based applications. Dehazing techniques turn a hazy image into a haze-free image [6], [9]. Typically, an atmospheric scattering model (ATSM) is used to do dehazing [6]. The creation of images in inclement weather is described by ATSM. The depth of the scene

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp135-146

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Dr. Alan Anwer Abdulla, Department of Information Technology, College of Commerce, University of Sulaimani, Department of Information Technology, University College of Goizha, Sulaimani, Iraq. E-mail: Alan.abdulla@univsul.edu.iq

Received: 28-09-2022 Accepted: 13-11-2022 Published: 01-12-2022

point affects the haze concentration according to ATSM [6]. Image dehazing can be broadly divided into three categories: 1 – Based on additional information, 2 – based on numerous images, and 3 – based on a single image. Early dehazing techniques are based on additional information [10], [11]. These techniques call for further details such as depth cues and degree of polarization. User engagement or other camera positioning procedures can supply this additional information. Therefore, real-time vision applications are not an appropriate for these approaches. For approaches based on multiple images, numerous images of a scene under various weather conditions obtained from differing degrees of polarization are required [12]–[16]. These techniques require additional hardware or resources; therefore, they are more expensive than single image dehazing, which has caught researchers' attention [6]. Strong priors and assumptions are the foundation of the majority of single image dehazing approaches. Only if the presumptions are accurate will these strategies work. Because these priors or assumptions were incorrect, the single image dehazing approaches were inaccurate. Typically, smoothing increases the transmission accuracy, which slows computing of the dehazing process. On the other hand, certain algorithms take atmospheric light into account [17], [18].

This paper concentrates on single-image dehazing which additional information is not required and also numerous images of different scene under various weather conditions are not required. It presents a new approach for selecting the haziest opaque regions of an image and using them to estimate the airlight. The rest of the paper contains the following sections: Section 2 presents the literature review. Section 3 explains the background. Section 4 describes the proposed approach. Section 5 shows the experimental results. Finally, the paper concludes in Section 6.

2. LITERATURE REVIEW

Usually, the unclarity of the images is produced due to the impact of weather conditions such as haze and fog. Haze/fog removal is considered a significant issue and hot topic by the researcher since the clarity of the degraded images is required for a variety of computer vision-based applications. Many techniques are introduced and existed for the haze/fog removal purposes in the area of DIP. The main competition in this research area is increasing the quality of the dehazed image; mainly, peak signal-to-noise-ratio (PSNR) and structural similarity index (SSIM) image quality measurements are used. This section reviews the most important and related existing works on image dehazing using DIP. In general, all the existing techniques either contributed to improve dark channel prior (DCP) and/or worked on proposing another technique to estimate the airlight and the transmission map. Furthermore, they used either PSNR, SSIM, or both to evaluate their performance.

This area of research was first explored by Narasimhan and Nayar in 2003, who offered a technique for reducing haze that made use of multiple images of the same location taken in various weather conditions [6]. They dealt with the issue of restoring contrast in images and videos that had been negatively affected by the atmosphere. This work discussed ways to identify depth discontinuities and determine a scene's structure using data from two images taken in various weather conditions. It demonstrated how to recover contrast from any image of the scene captured in inclement weather using either depth segmentation or the scene structure. This work did not use a specific dataset for the experimental result and it selected some images from the internet. The drawback of this approach is that it is dependent on weather variations to provide a number of images. In 2008, Raanan Fattal, suggested an image dehazing technique that requires only one input image [19]. Object surface shading and transmission signals are thought to be unrelated. The transmission map was calculated using independent component analysis. Later, a Markov random field was used to infer the color. For the experimental results, this work did not use a specific dataset, and the evaluation was performed on randomly selected images from the internet. Tarel and Hautiere in 2009, introduced a new image dehazing technique [20]. As reported in this work, since the ambiguity between the presence of fog and the items with low color saturation is resolved by assuming only small objects can have low color saturation, the ability to handle both color images and gray-level images is the achievement of this work. This work did not use a specific dataset for the experimental results and images are selected from different image datasets. Lu *et al.*, in 2015, proposed a powerful single-image dehazing technique [21]. Based on the color lines, airlight is calculated by applying a compensated filter to the white-balanced image, the highlight regions were eliminated as a pre-process for the airlight estimation. White-balanced image refers to the procedure of correcting colors to get objects that are white in reality to appear correctly white in your desired image by removing unnatural color casts. The airlight is estimated after the highlight regions have been subtracted from the image. The transmission map is then estimated using DCP. This work presented a semi-globally adaptive filter (SAF) to reduce the formation of gradient reversal artifacts on a rough transmission map. White-balanced image serves as the starting point for SAF's filtering procedure. To evaluate the performance of this work, the

AMOS-outdoor dataset was used. As authors reported, this technique achieved 15.2440 for PSNR and 0.7565 for SSIM. Salazar-Colores *et al.*, in 2019, introduced an image dehazing technique that significantly reduced the recurrent artifacts that are produced due to using the traditional DCP [22]. Both airlight and transmission map were estimated using DCP. This work provided a quick and an efficient way of altering the DCP computation, which greatly reduces the artifacts produced in the restored images when utilizing the standard DCP. It used the pixel-wise maximum operation to reduce the underestimated values in heterogeneous regions near from edges. However, the effect of the pixel-wise maximum operation values is essentially unaltered in homogenous regions, far from the edges, where the dark image neighbors are very comparable. For the experimental results, a dataset of 100 images has been created using Middlebury Stereo datasets. This technique obtained 18.50 of PSNR and 0.810 of SSIM. Dai *et al.*, in 2019, suggested a robust ATSM [23]. The actual image was breaking down into incident light as well as reflectance components and adding a noise term to the conventional model. For the airlight estimation, this work uses the same way as DCP which is picking 0.1% of brightest pixel in the dark channel of an image. Furthermore, it chooses eligible input image pixels as the candidate pixels because they have the same coordinates as the top 0.1% of the brightest dark channel pixels. This work also takes into account that the region with the greatest hazy opacity has a marginal three-channel difference. As a result, among the candidate pixels, the pixels with the smallest absolute change across the three channels are chosen to represent atmosphere light. To reduce over-enhancement in locations with thick haze, a compensation term with regard to transmission map is implemented after they estimate the transmission map using the DCP basis. For the performance evaluation, RESIDE SOTS (outdoor and indoor) and O-Haze datasets were used. Regarding to the RESIDE SOTS-outdoor dataset, 18.264 for PSNR and 0.855 for SSIM were obtained and for the RESIDE SOTS-indoor the obtained PSNR is 18.860 and SSIM is 0.831. Moreover, for the O-Haze dataset, 16.4 for PSNR and 0.75 for SSIM were gained. Gao *et al.*, in 2020, developed a non-local consistency assumption to eliminate the "halo" effects caused by standard image dehazing techniques and produce a haze-free image from a single hazy image [24]. When an image has been heavily edited, especially through the use of high dynamic range editing, a bright line known as a halo may form in places of high contrast on the image. This work dealt with each pixel separately instead of a block of pixels. It began by enhancing the technique for obtaining atmospheric light value and adapting their algorithm to a range of unique situations. Consequently, the brightness and saturation data were used for a single pixel to define a special energy function. The revised transmission map was then created by introducing propagation and random search into the image dehazing field. For the experimental results, 11 hazy images were selected from the internet and the obtained SSIM was reached 0.69. The limitation of this work is not perfect in some synthetic hazy image scenes which are caused by unaccurate airlight estimation. Zhang *et al.*, in 2020, developed a unique saliency-based and bright channel prior (BCP)-based single image dehazing technique [25]. A supper pixel-based atmospheric light estimation method was suggested to increase atmospheric light estimation accuracy in the stage of estimating atmospheric light. Furthermore, initially, the BCP model was suggested based on their observation to manage bright spots in the hazy images at the transmission map estimation stage. The automatic fusing of the DCP and BCP models is accomplished through a fusion-based transmission map estimation technique that is subsequently provided. In the refinement process, saliency analysis was applied to improve the rough transmission map. Middlebury Stereo dataset was used for testing this algorithm and the obtained PSNR and SSIM was 15.96 and 0.8287, respectively. In 2020, Yang and Wang. designed a new strategy to address DCP flaws [26]. This technique consists of two modules: Transmission map estimation and piece-wise function. In addition to acquiring a new dark channel map, a delicate function was used to replace the minimal filter operation. A nonlinear compression was then applied afterward to enhance accuracy and optimize the transmission map. When the ATSM used in conjunction, this technique can restore a clear image. For the experiment results, the RESIDE SOTS dataset was employed. Consequently, this technique achieved the PSNR of 17.021 and SSIM of 0.778.

Recently, Sun *et al.*, in 2021, suggested the new image dehazing technique [27]. The atmospheric light value is first calculated using the K-means clustering technique, which may successfully reduce the atmospheric light estimation inaccuracy brought on by the appearance of white objects in the image. Second, the transmission map is improved using the quick weighted guided filtering technique to eliminate of discontinuity and halo artifacts. The dehazing image's contrast and brightness are then adjusted using gamma correction and automatic contrast-enhancement methods. Middlebury dataset was utilized for the experimental outcome. Furthermore, for evaluating the proposed work, certain quality metrics were used such as information entropy, the rate of new visible edges, the mean of normalized gradients of visible edges, and average gradient. This technique effectively removes halo artifacts while restoring

clear images. In 2021, Raikwar and Tapaswi developed a method that utilized a difference channel (DCH) for estimate of the initial transmission map to nonlinearly translate the minimum channel of a hazy image into a minimum channel of a haze-free image [28]. This method employs a quad-tree subdivision-based method for the airlight estimation. It repeatedly divides an image into four rectangular parts. Based on the threshold, the brightest zone is selected as a region of atmospheric light. Contextual regularization enhances the smoothness of the initial transmission map. It has been demonstrated that estimation of the initial transmission map using DCH is more precise and reliable in variable haze concentration than existing methods. The suggested method can recover information from a distance while producing differing visual outcomes. However, regularization, a computationally slow approach is used to further smooth the initial transmission map obtained by the proposed method. The RESIDE SOTS and Dense-Haze datasets were used for the performance evaluation purposes. As reported in this work, for the RESIDE SOTS dataset, the obtained PSNR was 17.74 and the obtained SSIM was 0.83. Moreover, for the Dense-Haze dataset, the obtained PSNR was 12.26 and the obtained SSIM was 0.20.

Riaz *et al.*, in 2022, introduced a straightforward but efficient method of image restoration using multiple patches [29]. It fixed DCP's flaws and increased its computation speed for high resolution images. For the airlight, this work uses the same techniques as DCP. The smallest number of patches of various sizes is used to estimate a coarse transmission map. The transmission map is then improved using a cascaded rapid guided filter. This work provides the advantage of very little performance reduction for a high resolution image by introducing an effective scaling technique for the transmission map estimation. The standard Middlebury stereo vision dataset was utilized for the performance evaluation. Furthermore, this work reached a SSIM value of 0.9689.

The proposed approach presented in this paper focuses on single-image dehazing and it improves the DCP by enhancing the airlight estimation, details are discussed in the next sections.

3. BACKGROUND

This section concerns with the explanation of the atmosphere scattering model (ATSM) to understand how the hazy image is formed. In addition, the DCP is also discussed as one of the most mechanism used by the majority of researchers as a base for improving it or develop their technique for image dehazing.

3.1. ATSM

The ATSM was introduced by McCartney in 1976, aims to explain how hazy images are formed [30]. Later, the ATSM was significantly improved by Narasimhan and Nayar [6]. An imaging model of a hazy scene, as illustrated in Fig. 1, essentially consists of two factors under the principle of atmospheric scattering: (1) The technique of attenuating the light that is reflected from an object's surface onto a camera is the first factor, (2) the second factor is how the airlight is dispersed as it approaches the camera. Theoretically, hazy and blurry images are based on both components [31].

Consequently, the scattering model to represent hazy images in the field of computer vision can be expressed as:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (1)$$

Where x is the distance coordinate, $I(x)$ denotes an image with the haze, $J(x)$ denotes an image without haze, A denotes an atmospheric light, and $t(x)$ is the medium's transmission rate which is also known as transmission map. To recover $J(x)$ from $I(x)$, image dehazing is used. The deterioration model has several unidentified parameters, which creates an imprecise problem. $J(x)$ can only be reconstituted from $I(x)$ after estimating the parameters A and $t(x)$.

3.2. DCP

The DCP is based on the observation that most non-sky areas have at least one color channel with very low intensity at certain pixels in haze-free outdoor images [12]. In other words, the minimum intensity of the patch/block should be quite low. Formally, for an image J, the DCP can expressed as:

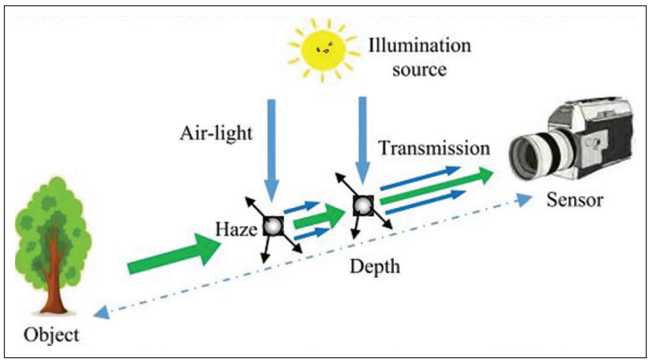


Fig. 1. Imaging model of hazy scene [31].

$$J_{dark}(x) = \min_{c \in \{r,g,b\}} \left(\min_{y \in \Omega(x)} J^c(y) \right) \quad (2)$$

Where $_{\Omega}(x)$ is a local patch centered at x, y is one pixel in the local patch $_{\Omega}(x)$, J^c is a color channel of J. According to the observation, if J is a clear outdoor image, the intensity of J_{dark} is low and typically zero, with the exception of the sky region. The above statistical finding or information is referred to as the DCP, and J_{dark} referred to as the dark channel of J. According to [12], the following three reasons contribute to low intensities in the dark channel:

3.2.1. Shadows

For instance, the shadows cast by vehicles, buildings, and the interiors of windows in cityscape images, or by leaves, rocks, and trees in landscape images.

3.2.2. Bright surfaces or items

For instance, any object such as green grass, trees, plants, red or yellow flowers, or blue water surfaces lacking color in any color channel will produce low values in the dark channel.

3.2.3. Dark items or regions

The black channels of some images are incredibly dark because the natural outside images are typically colorful and full of shadows. For instance, stone or a black tree trunk.

In general, the DCP consists of two stages [12]. First stage is the airlight estimation which is selecting the top 0.1% of pixels in J_{dark} with the highest brightness found, and then the maximum value of the pixels that match the pixels in the original image is chosen as the atmospheric light. The second stage is the transmission map estimation, which is according to [12] can be expressed as:

$$t(x) = 1 - \min_{c \in \{r,g,b\}} \left(\min_{y \in \Omega(x)} \left(\frac{I^c(y)}{A^c} \right) \right) \quad (3)$$

I_c and A_c represent input hazy image and airlight in color channel c, respectively. In practice, to preserve the sense of depth in the image, a correction factor ω $(0<\omega\leq 1)$ is added to keep the partial haze. Then, Equation (3) can be rewritten as follows:

$$t(x) = 1 - \left(\omega * \min_{c \in \{r,g,b\}} \left(\min_{y \in \Omega(x)} \left(\frac{I^c(y)}{A^c} \right) \right) \right) \quad (4)$$

Since the regional transmission map is assumed to be constant, the block effect often exists in the transmission map. To further improve $t(x)$, soft-matting [12] or guided filtering [16] are applied. The final scene radiance can be recovered using Equation (5) under the ATSM once the transmission map $t(x)$ and the atmospheric light A have been acquired:

$$J(x) = \frac{I(x) - A}{t(x)} + A \quad (5)$$

4. PROPOSED APPROACH

Based on DCP, two essential components of single image dehazing are airlight estimation and transmission map estimation [12]. In DCP, before transmission map estimation, the input hazy image must be first normalized by the estimated airlight. Airlight estimation is crucial for recovering haze-free scene radiance. Consequently, estimating the airlight improperly leads to an erroneous transmission map estimation and incorrect scene radiance recovery. Based on the DCP strategy, the dark channel of the image can be calculated using Equation (2). Then, 1% of the brightest pixel in the dark channel is selected and then calculating the average of the matching pixel in the input hazy image is used as an airlight. This estimation fails when there are white items in the image and when these white things are chosen as the scene's most opaque haze region. In this study, the proposed approach improves the airlight estimation that uses the HSV color space. The brief detailed of the proposed approach is presented in the following subsection.

4.1. Airlight Estimation

Due to the impact of haze, certain regions of the hazy image are caused to have high brightness and low saturation. For a haze-free region or light haze in a hazy image, the scene's saturation is rather high, its brightness is moderate, and the difference between brightness and saturation is almost near to zero [32]. However, in [32], the authors discovered that for the places with moderate haze, the saturation of the region drastically falls while the color of the scene fades due to the haze, and the brightness increases at the same time causing the large value of the difference. It is more challenging for human eyes to distinguish the scene's natural color in areas with strong haze, and the difference is even greater. According to [32], it appears that the three characteristics (brightness, saturation, and difference) are likely to change frequently in a single hazy image. Consequently, in this study, the HSV color spaces is utilize to pick the region of the hazy image that contains the haziest and then use that region to estimate the airlight. The steps of the proposed approach for the airlight estimation are as follows:

1. The input hazy image I is converted from RGB to HSV color space producing I'

2. The image I is divided into blocks of size (32 × 32) pixels
3. For each block, the difference between brightness V and saturation S is calculated, $D = | V - S |$
4. The block that has a maximum difference value, i.e. maximum D, is selected as the haziest opaque region
5. The dark channel for the selected block needs to be produced using Equation (2)
6. From the produced dark channel, 1% of the brightest pixels in the block (i.e., 0.01 × 32 × 32 = 10 pixels) are selected
7. Based on the location of the selected 10 brightest pixels, select 10 pixels of the same location in the original block of the input hazy image, and then calculate the average A of them for each channel separately producing A_R, A_G, A_B
8. Finally, A_R, A_G, and A_B are considered the values of airlight.

Fig. 2 illustrates the block diagram of proposed airlight estimation.

The advantage of the proposed airlight estimation over the airlight estimation in DCP is that DCP can fail to select the haziest opaque region by the influence of white object. While in the proposed airlight estimation, the region with haziest can be selected, and hence, the airlight can be estimated from that region properly and more accurately. Furthermore, instead of calculating the dark channel for the entire image, the proposed approach calculates the dark channel of only one block of the input image. This leads to reduce the time consumption, since calculating dark channel for the entire image is time consuming.

4.2. Transmission Map Estimation

As previously mentioned, the DCP is divided into two essential components. The first is airlight estimation, which is briefly detailed in Section 4.1. Estimating the transmission map is the second component, which can be carried out using Equation (4). Due to atmospheric particles, which are presented even on clear days. Therefore, the haze is still presented when human looks at far-off objects. If the haze is completely removed, the image could appear unnatural and the sense of depth might disappear. Therefore, the authors in [12] alternatively introduce a constant parameter ω ($0 < \omega < 1$) to maintain a very little level of haze for distant objects. The value of ω depends on the application. The majority of works set it at 0.95. For all of the results described in this study, also 0.95 were used. Moreover, the patch size of (15 × 15) was used, the same as that used in the DCP. The obtained transmission map contains block effects since a patch's transmission is not always constant, The transmission map is improved using the guided filter [16].

4.3. Scene Radiance Recovery

On the bases of Equation (5), the haze-free scene can be reconstructed once the airlight and transmission map have been estimated. Fig. 3 illustrates the block diagram of the scene radiance recovery, that is, dehazing technique.

5. EXPERIMENTAL RESULTS

This section concerns with the performance evaluation of the proposed image dehazing approach. First, it contains details about the dataset that was used in the experimental results. Second, to evaluate the influence of the proposed approach

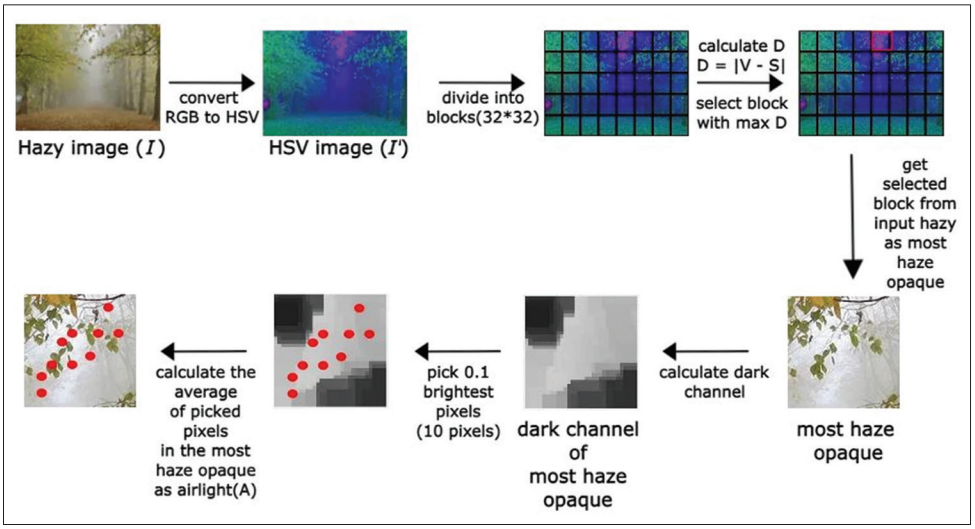


Fig. 2. Block diagram of the proposed airlight estimation.

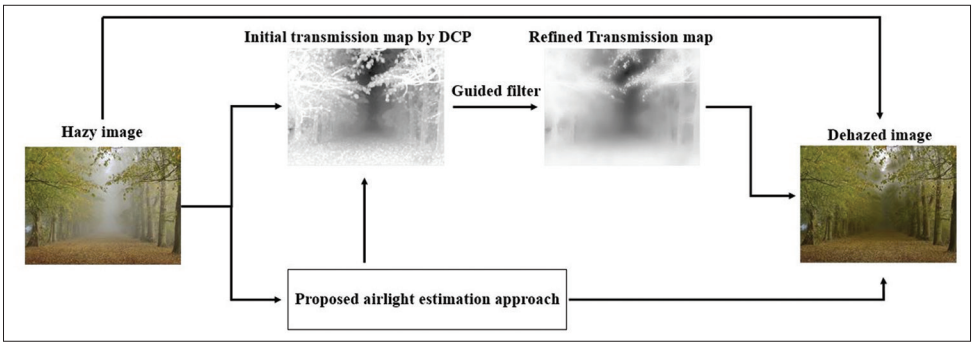


Fig. 3. Block diagram of the Dehazing technique.

in terms of image quality, objectively as well as subjectively, extensive experiments are carried out. Finally, the results of the proposed approach are compared to the results of the most recent relevant approaches.

6. EXPERIMENTAL ENVIRONMENT

To evaluate the effectiveness of the suggested strategy, experiments are conducted under the Intel i7-6600U 2.8 GHz CPU and 8 GB RAM, using Matlab 2018. The proposed approach uses RESIDE SOTS dataset which contains 500 images for each outdoor and indoor scene [33]. Figs. 4 and 5 show some image examples for each indoor and outdoor dataset, respectively.

6.1. Objectively Assessment

Image quality evaluation is crucial in image analysis systems to analyze techniques and assess their performance. Consequently, it is essential to analyze the experimental findings objectively. Furthermore, two common objective evaluation methods are used, which are:

6.1.1. SSIM

SSIM is one of the most common image quality measurement. For employing this measurement, two images from the same acquired image must be considered, the original image and the processed image. SSIM can be calculated using the following Equation [34].

$$SSIM\ (x,y) = [\,l\,(x,y)\,]\,\alpha\ .\left[c\,(x,y)\right]\beta.[\,s\,(x,y)\,]\,\vartheta \quad (6)$$

$l(x,y)$ is a luminance comparison function, $c(x,y)$ refers to contrast comparison function, and $s(x,y)$ is structure comparison function. Moreover, x and y are two images to be compared. In addition, $\alpha > 0$, $\beta > 0$, $\vartheta > 0$ denote the relative importance of each of the metrics.

6.1.2. PSNR

The PSNR ratio measures how much noise can degrade an image's representational quality in comparison to its highest achievable power. Calculating the PSNR, it is necessary to compare that image to an ideal clean image with the maximum possible power. The Equation 7 can be used to compute PSNR [35].

$$PSNR = 10\ log\ 10\left(\frac{(L-1)^2}{MSE}\right) \quad (7)$$

Where, L is the value of maximum possible intensity levels. MSE is the mean squared error and it is defined as:

$$MSE = \frac{1}{m\ n}\sum_{i=0}^{M-1}\ \sum_{j=0}^{N-1}(\,I\,(i,j)-R\,(i,j)\,)^2 \quad (8)$$

Where:
I: represents the matrix data of the original image
R: represents the matrix data of the reconstructed/degraded image
m: represents the number of rows of pixels
i: represents the index of that row of the image
n: represents the number of columns of pixels
j: represents the index of that column of the image.

The obtained results are compared to the results of four other existing techniques such as [12], [23], [26], and [28], Table 1.

Table 1 presents the average value of SSIM and PSNR for the all images in the tested dataset. Table 1 makes it abundantly clear that the proposed approach outperformed the existing approaches. Moreover, the PSNR and SSIM of the proposed approach are significantly increased in comparison to other existing techniques.

The reason behind selecting reference [12] is, this work is considered as one of the most successful and earliest work in the area of single image dehazing. Moreover, the majority

Fig. 4. Example of outdoor images.

Fig. 5. Example of indoor images.

Fig. 6. Reconstructed outdoor images: (a) Hazy images, (b) Reconstructed by [12], (c) Reconstructed by [23], (d) Reconstructed by [26], (e) Reconstructed by [28], (f) Reconstructed by proposed approach.

of the existing works are based on [12] that known as DCP in the literature. They either present their contribution by making improvements to DCP or by defining new techniques for making DCP work better. In addition, the reasons behind selecting the references [23], [28], and [26] are because these works are recently published, their contribution is significant, and they were published in the well-known and high-quality journals.

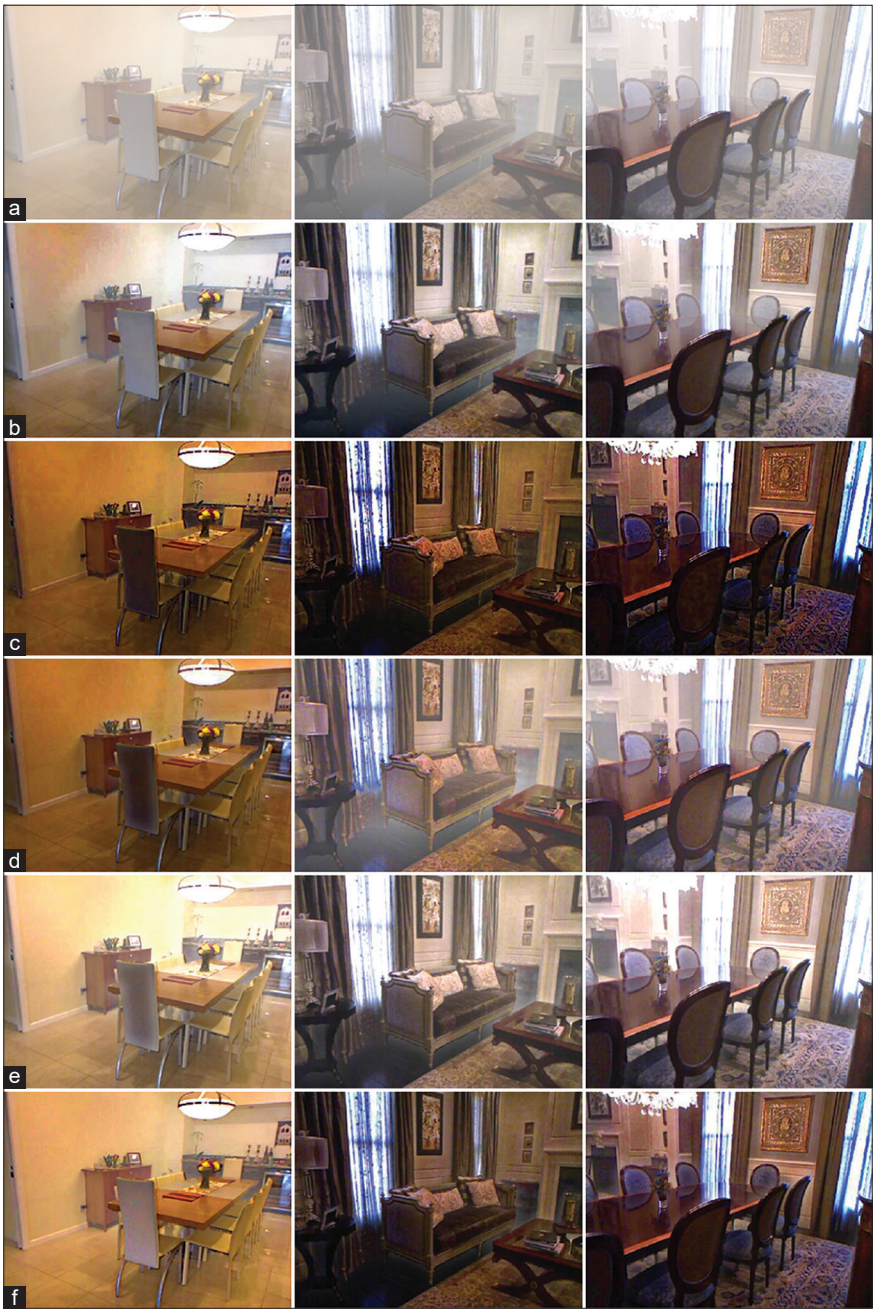


Fig. 7. Reconstructed Indoor Images: (a) Hazy images, (b) Reconstructed by [12], (c) Reconstructed by [23], (d) Reconstructed by [26], (e) Reconstructed by [28], (f) Reconstructed by proposed approach.

6.2. Subjectively Assessment

Figs. 6 and 7 illustrate the reconstructed images using the proposed image dehazing approach and other existing approaches for both SOTS-outdoor and SOTS-indoor, respectively.

From Fig. 6, one can notice that the resulted images of the proposed approach clearer and the haze removed properly compared to other tested techniques. In addition, it can be noticed that results of [12] and [28] are still hazy and the haze not removed completely, while results of [23] and [26] are over dehazed.

From Fig. 7, it is quite obvious that in the resulted images of the proposed approach, the haze is completely removed and the colors of the scenes are rendered naturally. In contrast,

TABLE 1: Objectively performance evaluation of the proposed approach and the existing approaches

Approaches	SOTS-Outdoor		SOTS-Indoor	
	PSNR	SSIM	PSNR	SSIM
[12]	16.45	0.86	19.33	0.87
[23]	18.2	0.88	18.8	0.83
[28]	17.19	0.86	18.29	0.80
[26]	18.29	0.86	16.07	0.80
Proposed approach	18.37	0.90	20.59	0.89

the resulted images of [23] are mostly over dehazed and for the techniques [26], [28], and [12] the haze is not removed completely.

7. CONCLUSION

Images acquired under hazy environment require processing for improving their contrast and color fidelity. Haze removal or dehazing is a significant pre-processing stage in the area of computer vision and video applications. Many techniques have been proposed in the literature for dehazing outdoor/indoor images. This study is presented an approach to enhance atmospheric airlight by exploiting HSV color space. The proposed approach can discover the most haze opaque region in the input hazy image by finding the difference between brightness and saturation of each region of the input hazy image. Consequently, from the selected region, the airlight is estimated. Regarding to the transmission map, the proposed approach uses the traditional DCP technique. In other words, the proposed approach focused on improving the airlight. The proposed approach is implemented on RESIDE SOTS dataset for both outdoor and indoor images. The performance of the proposed approach is assessed objectively and subjectively. Regarding to the objectively evaluation, the proposed approach achieved the PSNR and SSIM of 18.37 and 0.90 for the outdoor images, respectively. For the indoor images, it achieved the PSNR and SSIM of 20.59 and 0.89, respectively. The obtained results are compared to the results of other existing image dehazing techniques in terms of PSNR and SSIM and it outperformed existing techniques. In terms of subjectively evaluation, the proposed approach again outperformed the existing techniques. The future direction of this research will concentrate with strengthening and improving transmission map estimation to obtain/reconstruct the better dehazed image quality.

REFERENCES

[1] A. A. Abdulla and M. W. Ahmed. “An improved image quality algorithm for exemplar-based image inpainting”. *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13143-13156, 2021.

[2] S. F. Salih and A. A. Abdulla. “An improved content based image retrieval technique by exploiting bi-layer concept”. *UHD Journal of Science and Technology (UHDJST)*, vol. 5, no. 1, pp. 1-12, 2021.

[3] J. Y. Kim, L. S. Kim and S. H. Hwang. “An advanced contrast enhancement using partially overlapped sub-block histogram equalization”. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 475-484, 2001.

[4] M. J. Seow and V. K. Asari. “Ratio rule and homomorphic filter for enhancement of digital colour image”. *Neurocomputing*, vol. 69, no. 7-9, pp. 954-958, 2006.

[5] Z. Rahman, D. J. Jobson, and G. A. Woodell. “Retinex processing for automatic image enhancement”. *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 100-110, 2004.

[6] S. G. Narasimhan and S. K. Nayar. “Contrast restoration of weather degraded images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 713-724, 2003.

[7] S. G. Narasimhan and S. K. Nayar. “Removing weather effects from monochrome images”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* CVPR, vol. 2, p. 2, 2001.

[8] E. Namer, S. Shwartz and Y. Y. Schechner. “Skyless polarimetric calibration and visibility enhancement”. *Optics Express*, vol. 17, no. 2, pp. 472-493, 2009.

[9] F. Liu, L. Cao, X. Shao, P. Han and X. Bin. “Polarimetric dehazing utilizing spatial frequency segregation of images”. *Applied Optics*, vol. 54, no. 27, pp. 8116-8122, 2015.

[10] S. G. Narasimhan and S. K. Nayar. “Interactive (de) weathering of an image using physical models”. In: *IEEE Workshop on color and photometric Methods in computer Vision*, vol. 6, no. 6.4, p. 1.

[11] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele and D. Lischinsk. “Deep photo: Model-based photograph enhancement and viewing.” *ACM Transactions on Graphics*, vol. 27, no. 5, pp. 1-10, 2008.

[12] K. He, J. Sun and X. Tang. “Single image haze removal using dark channel prior”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341-2353, 2010.

[13] J. Long, Z. Shi and W. Tang. “Fast haze removal for a single remote sensing image using dark channel prior”. *2012 International Conference on Computer Vision in Remote Sensing*, pp. 132-135, 2012.

[14] G. Meng, Y. Wang, J. Duan, S. Xiang and C. Pan. “Efficient image dehazing with boundary constraint and contextual regularization”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 617-624, 2013.

[15] J. B. Wang, N. He, L. L. Zhang and K. Lu. “Single image dehazing with a physical model and dark channel prior”. *Neurocomputing*, vol. 149, pp. 718-728, 2015.

[16] K. He, J. Sun and X. Tang. “Guided image filtering”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397-1409, 2012.

[17] A. Levin, D. Lischinski and Y. Weiss. “A closed-form solution to natural image matting”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228-242, 2007.

[18] D. Berman and S. Avidan. “Non-local image dehazing”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1674-1682, 2016.

[19] R. J. Fattal. “Single image dehazing”. *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 1-9, 2008.

[20] J. P. Tarel and N. Hautiere. “Fast visibility restoration from a single color or gray level image”. In: *2009 IEEE 12th international conference on computer vision*, pp. 2201-2208, 2009.

[21] H. Lu, Y. Li, S. Nakashima and S. Serikawa. “Single image dehazing through improved atmospheric light estimation”. *Multimedia Tools and Applications*, vol. 75, no. 24, pp. 17081-17096, 2016.

[22] S. Salazar-Colores, J. M. Ramos-Arreguín, J. C. Pedraza-Ortega and J. Rodríguez-Reséndiz. “Efficient single image dehazing by modifying the dark channel prior”. *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p.66, 2019.

[23] C. Dai, M. Lin, X. Wu and D. Zhang. “Single hazy image restoration using robust atmospheric scattering model”. *Signal Processing*, vol. 166, p. 107257, 2020.

[24] Y. Gao, Y. Zhang, H. Li and W. Zhang. “Single image dehazing based on single pixel energy minimization”. *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5111-5129, 2021.

[25] L. Zhang, S. Wang and X. Wang. “Single image dehazing based on bright channel prior model and saliency analysis strategy”. *IET Image Processing*, vol. 15, no. 5, pp. 1023-1031, 2021.

[26] Y. Yang and Z. Wang. “Haze removal: Push DCP at the edge”. *IEEE Signal Processing Letters*, vol. 27, pp. 1405-1409, 2020.

[27] F. Sun, S. Wang, G. Zhao and M. Chen. “Single-image dehazing based on dark channel prior and fast weighted guided filtering”. *Journal of Electronic Imaging*, vol. 30, no. 2, p. 021005, 2021.

[28] S. C. Raikwar, S. Tapaswi. “Estimation of minimum color channel using difference channel in single image Dehazing”. *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 31837-31863, 2021.

[29] S. Riaz, M. W. Anwar, I. Riaz, H. W. Kim, Y. Nam and M. A. Khan. “Multiscale image dehazing and restoration: An application for visual surveillance”. *Computers, Materials and Continua*, vol. 70, pp. 1-17, 2021.

[30] E. J. McCartney. “Optics of the atmosphere: Scattering by molecules and particles”. *Physics Bulletin*, p. 421, 1976.

[31] W. Wang, F. Chang, T. Ji and X. Wu. “A fast single-image dehazing method based on a physical model and gray projection”. *IEEE Access*, vol. 6, pp. 5641-5653, 2018.

[32] Q. Zhu, J. Mai and L. Shao. “A fast single image haze removal algorithm using color attenuation prior”. *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522-3533, 2015.

[33] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng and Z. Wang. “Benchmarking single-image dehazing and beyond”. *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492-505, 2019.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh and E. Simoncelli. “Image quality assessment: From error visibility to structural similarity”. *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[35] A. Hore and D. Ziou. “Image quality metrics: PSNR vs. SSIM”. In: *2010 20th International Conference on Pattern Recognition*, pp. 2366-2369, 2010.

SURVEY UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Real-Time Twitter Data Analysis: A Survey

Hakar Mohammed Rasul, Alaa Khalil Jumaa

Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq

ABSTRACT

Internet users are used to a steady stream of facts in the contemporary world. Numerous social media platforms, including Twitter, Facebook, and Quora, are plagued with spam accounts, posing a significant problem. These accounts are created to trick unwary real users into clicking on dangerous links or to continue publishing repetitious messages using automated software. This may significantly affect the user experiences on these websites. Effective methods for detecting certain types of spam have been intensively researched and developed. Effectively resolving this issue might be aided by doing sentiment analysis on these postings. Hence, this research provides a background study on Twitter data analysis, and surveys existing papers on Twitter sentiment analysis and fake account detection and classification. The investigation is restricted to the identification of social bots on the Twitter social media network. It examines the methodologies, classifiers, and detection accuracies of the several detection strategies now in use.

Index Terms: Twitter, Data Analysis, Twitter Streaming Application Programming Interface, Sentiment Analysis, Bot Detection

1. INTRODUCTION

The availability of data has increased dramatically since the Big Data era began, and it is predicted that this trend will continue in the years to come. A thorough research is being done to make appropriate use of this knowledge. Big data and big data analytics have created opportunities for businesses and scholars that were previously unthinkable. Research in artificial intelligence on how to leverage readily available data is producing fascinating and important results. There are several sources of big data, and one of the most well-known ones is social networks, including Twitter.

Twitter is a microblogging social network that enables users to post short messages (up to 280 characters) called tweets. Users may interact with one another on Twitter by responding to tweets, referencing other users in their tweets, or retweeting another user's message. Users can also follow each other to keep up with what other people are saying on Twitter. All registered users have access to the social network's services through a web page, mobile apps, and an application programming interface (API). The latter method of access has produced an ecosystem of applications that enhance the user's experience of information consumption and aggregation. However, this has aided in the development of systems for account management and automated tweet publishing [1]. Fully-automated accounts are called bots. They can retweet exciting and relevant material for specific communities or aggregate tweets about a topic.

One area that requires attention is bot detection analysis. Since, around 48 million Twitter accounts have been maintained by automated programs dubbed bots, accounting for up to 15% of all Twitter accounts [2]. Certain bots are helpful for a numerous task, including automatically publishing news and academic articles and aiding in emergency circumstances. Nonetheless, Twitter bots have been used for malicious

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp147-155 | E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq. E-mail: Hakar.mohammed.r@spu.edu.iq

Received: 17-06-2022 Accepted: 11-11-2022 Published: 21-12-2022

purposes, such as spreading malware or manipulating public opinion on a certain subject.

Bot identification software is predicated on the premise that the behavior of a human account is distinct from that of a bot. To quantify these discrepancies, representative factors including the statistical distribution of the terms used in tweets, the frequency of daily posts, and the number of individuals who followed the user may be employed [3]. Apache Spark data analysis on Twitter will be required to do that. As a result, the portion that follows in this essay will examine similar efforts on Twitter data analysis employing Apache Spark and bot identification, as well as the available tools.

2. BACKGROUND INFORMATION

2.1. Twitter

Twitter is a microblogging and social networking website that enables users to post and receive 280-character messages called "tweets." Registered users may send tweets and follow other users. Unregistered users may browse public tweets on Twitter without having an account [4].

Over 300 million individuals use Twitter on a regular basis. More than 500 million tweets each day are sent in 33 different languages [5]. One of Twitter's best benefits is the capacity for communication and sharing with other users. By sharing links, pictures, and videos with their followers, people and businesses may interact with them [6]. This section explains some of Twitter features:

1. Follow: To follow someone on Twitter, you must subscribe to their tweets or site updates. Another Twitter user who has followed you is referred to as a "Follower." Other Twitter users you've decided to follow on the platform are referred to as "following." [7]
2. @: In tweets, the @ symbol is used to identify usernames. The @ sign before a username (like @HakarRasul) creates a connection to that Twitter user's profile [8].
3. Reply: A tweet in response to a tweet from another person. To answer to a tweet, users often click the "reply" box or icon adjacent to it. @username is always the first character in a reply [9].
4. Retweet: The act of forwarding another user's tweet is denoted as "retweeting." In essence, you are sharing another user's tweet in your profile while properly acknowledging the message's original writer [10].
5. Mention: This term refers to tweets that contain a username. @replies are a type of mention as well [11].
6. Hashtag: The # symbol is used in tweets to denote topics or keywords. Hashtags are limited to letters and numbers (no punctuation). Other Twitter users may use a hashtag you tweet to search for it. Any Twitter user may generate a hashtag at any moment [12].
7. Direct Messages: These Tweets, sometimes referred to as direct messages or simply "messages," are confidential between the transmitter and recipient. When you start a tweet with "d username" to identify the recipient, the tweet becomes a direct message (DM). You should be following someone to send them a Direct Message [13].
8. Trends: A subject recognized by Twitter's algorithm as among the hottest subjects on the network right now [14].
9. Favorites: To add a Tweet to your favorites, click the yellow icon next to the tweet. Tweets you've favorite will stay in your list until you delete them [15].

2.2. Twitter Streaming API

The Twitter API now includes a Streaming API in addition to two separate REST APIs. The streaming API provides real-time access to Tweets that were sampled and filtered. The API is HTTP-based, with data accessible through GET, POST, and DELETE requests. The streaming API allows you to access subsets of public status descriptions, such as answers and mentions from public accounts, in near-real time. Protected users' status descriptions and direct messages are no longer viewable. The streaming API may filter status descriptions based on quality criteria, which are influenced, in addition to, by frequent and repeated status updates [16].

The API requires a valid Twitter account and employs simple HTTP authentication. Data may be obtained in both XML and the shorter JSON format. The parsing of JSON data got from the streaming API is straightforward: Each object is delivered on a separate line, with a carriage return at the conclusion [17].

Twitter streaming data allow every user to learn about what is going on in the globe at any given moment. The Twitter streaming API provides access to a huge quantity of tweets in real time [18].

A Python package called Twitter4j is available to access the streaming API and download Twitter data to analyze data from the Twitter API. This data has been filtered using a list of provided keywords. This research will use Apache Spark, a distributed data processing system with many workers, and master nodes. This cluster can manage millions of records and is scalable. Map reduction on Spark might be used to filter out the massive amount of data. For each tweet in the data, a JSON object will be included in the input file. On the Spark frame structure, this file will be uploaded. The mapper classifies all files in the directory according to the filter specified once

the Spark frame structure has been duplicated and distributed across several nodes. These cleaned tweets will go through data mining techniques, allowing for a one-to-one analysis of data that will be useful for making difficult judgments [19].

2.3. Analysis Process on Twitter Data

There are several steps to be performed to analyze Twitter data. Fig. 1 illustrates the phases of analyzing Twitter data.

2.3.1. Dataset collection

To gather real-time data, an application must be developed that uses the Twitter API to capture the information of people who recently tweeted about the issue and construct a user-based feature set data frame [21].

2.3.2. Processing tweets

This stage involves removing unnecessary material from tweets in the style of regular expressions [22].

2.3.3. Feature selection

Here, some of features should be considered, such as the user-based and content-based. These features have to be selected to enhance the detection and classification process [21].

2.3.4. Classification

In this step, the user must be checked in real-time whether it is a bot or human, which may be accomplished by training and testing the proposed model. The proposed model can be build using one of the machine learning algorithms. After applying the machine learning algorithm on a preprocessed, existed, and labeled dataset, a model can be created. Then, this model can be used to predict if the streamed Twitter that we got from Twitter is human or bot [22].

3. METHODOLOGY

This section will provide the clarification of the searching, filtering, and stages that were employed throughout this paper's research stage.

3.1. Research Sections

In Section I, questions like (What is big data, what is Twitter, and what is the connection between Twitter and big data?) has been answered. Then, Section II explained Twitter and its important elements; Twitter API and its use; and the phases of Twitter data processing. Section III gives a methodology about how this paper been organized and the methods that have been used to gathered information. Section IV provides a survey methodology and has been divided into two parts survey of articles about sentiment analysis and survey of articles about bot detection and classifications.

Fig. 1. Twitter data analysis process [20].

3.2. Search Query

This paper aims to summarize the current state of the real-time Twitter data analysis topic and discuss the findings presented in recent research papers. Hence, those keywords have been used.

("Twitter data") AND ("Real-Time OR "Bots") AND ("Sentiment analysis" OR "Bot Classification" OR "Data Extraction" OR "Preprocessing" OR "Text-mining" OR "web-Mining") AND ("Challenges" OR "Problems" OR "Patterns").

3.3. Selection of Sources

Google Scholar and Elsevier have been used for applying the search queries and the databases that have been considered were IEEEXplore Digital Library, SpringerLink Journal, Elsevier, and Science Direct.

3.3.1. Selection phases

Each article that has been chosen to be used in this paper has been gone through these processes:

The first phase of article selection is applying the search queries. Then, select only the articles have been published between 2016 and 2021. After that, the title of the research and the list of index terms had been considering to see if it includes the keyword "Twitter, Data Analysis."

The next step was reading the abstract and the conclusion of the paper, and selecting the paper according to its abstract and conclusion then the relatively of its body to them. Finally, the last phase was considering the journal's indexing and if they are peer reviewed or not.

4. SURVEY METHODOLOGY

When coming to Twitter data analysis, there are various types of analysis that might be done on the collected data such as sentiment analysis, tweets classification, and fake tweets detection. Hence, this survey will be categorized into two sections (A) sentiment analysis and (B) tweets classification and bot detection. Table 1 shows list the studies that have been surveyed in this section.

TABLE 1: List of studies that have been reviewed in Section IV

Title (s)	Author (s)	Technique (s)	Result (s)	Year
Sentiment Analysis				
"Sentiment analysis and classification of Indian farmers' protest using twitter data"	Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi	Bag of Words and TF-IDF	Bag of Words was more effective than TF-IDF.	2022
"An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis"	T. Swathi, N. Kasiviswanath, A. Ananda Rao	TLBO-LSTM	Precision: 0.95, Recall 0.85, Accuracy: 0.94, F1-score 0.90	2022
"Twitter Sentiment Analysis during COVID-19 Outbreak"	Akash Dutt Dubey	NRC Emotion Lexicon	The majority of individuals around the globe are optimistic.	2020
"Detection of Fake Tweets Using Sentiment Analysis"	C. Monica, N. Nagarathna	Rule-based prediction	Accuracy: 0.97, F1-score: 0.73, Precision: 1.00, Recall: 0.97	2020
"Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers"	Gonzalo A.Ruz, Pablo A. Henríquez, Aldo Mascareño	Bayes factor	Accuracy: 0.85, Precision: 0.92, Recall: 0.77, F1-score: 0.82	2020
"Twitter Sentiment Analysis Based on Ordinal Regression"	Shihab Elbagir Saad, Jing Yang	Multinomial logistic regression (SoftMax), Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF)	Accuracy: 0.91, F1-score: 0.85 using Decision Tree.	2019
Classification and bot detection				
"The Rise of Social Bots," "Online human-bot interactions: Detection, estimation, and characterization," "Deep Neural Networks for Bot Detection," "Evolution of bot and human behavior during elections," "Measuring bot and human behavioral dynamics"	Emilio Ferrara *et al*.	Session features	Accuracy: 0.97	2016 2017 2018 2019 2020
"A deep learning model for Twitter spam detection"	Zulfikar Alom. Barbara Carminati, Elena Ferrari	Deep learning	Accuracy: 0.99, Recall: 0.98, F1-score: 0.93	2020
Classification and bot detection				
"Twitter Bot Detection Using Bidirectional Long Short-Term Memory Neural Networks and Word Embeddings"	Feng Wei, Uyen Trang Nguyen	Recurrent neural networks, specifically bidirectional Long Short-term Memory (BiLSTM)	Accuracy: 0.92, Precision: 1.00, Recall: 0.85, F1-score: 0.92	2019
"Social Network Polluting Contents Detection through Deep Learning Techniques"	Fabio Martinelli, Francesco Mercaldo, Antonella Santone	Combination of word embedding and deep learning	Precision: 0.79, Recall: 0.73, F1-score: 0.76	2019
"DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks"	Qingyuan Gong, Yang Chen, Xinlei He, Zhou Zhuang, Tianyi Wang, Hong Huang, Xin Wang, Xiaoming Fu	long short-term memory (LSTM) neural network	Precision: 0.95, Recall: 0.97, F1-score: 0.96	2018
"Measuring bot and human behavioral dynamics"	Iacopo Pozzana, Emilio Ferrara	Extra Trees (ET), DT, Random Forests (RF), Adaptive Boosting (AB), and KNN	ET and RF had the greatest cross-validated average performance 0.86	2018
"Deep neural networks for bot detection"	Sneha Kudugunta, Emilio Ferrara	Deep neural network based on contextual long short-term memory (LSTM)	Accuracy: 0.96, Precision: 0.96, Recall: 0.96, F1-score: 0.96	2018
"Classification of Twitter Accounts into Automated Agents and Human Users"	Zafar Gilani, Ekaterina Kochmar, Jon Crowcroft	Random Forests classifier	Accuracy: 0.86, Precision: 0.85, Recall: 0.82, F1-score: 0.83	2017

(*Contd...*)

TABLE 1: (*Continued*)

Title (s)	Author (s)	Technique (s)	Result (s)	Year
"Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?"	Zi Chu, Steven Gianvecchio, Haining Wang, Sushil Jajodia	Bayesian classification	overall system accuracy: 96.0	2012
"Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach"	Alex Hai Wang	Decision Tree (DT), Neural Network (NN), Support Vector Machines (SVM), Naive Bayesian (NB), and k-Nearest Neighbors, are used to detect spam bots (KNN)	Accuracy: 0.91, Precision: 0.91, Recall: 0.91, F1-score: 0.91 using NB	2010

4.1 Sentiment Analysis on Twitter

A computer finding the mood of a word, phrase, or tweet is quite challenging. To ascertain the polarity of the words and perform sentiment analysis, human participation is required. Since it is used to evaluate people's sentiments, views, and emotions, this form of analysis is sometimes referred to as "opinion mining." It is done by evaluating each word's attitude and classifying it as either positive, negative, or neutral. In addition, there are other ways to do sentiment analysis, including by employing a lexicon, machine learning, deep learning, or a combination of machine learning and lexicon-based approaches. In the lines that follow, recent studies on sentiment analysis on Twitter will be reviewed.

Neogi *et al.* [23] acquired data from the microblogging website Twitter on farmer protests to comprehend the global views shared by the public. They categorized and analyzed the attitudes based on over 20,000 tweets about the demonstration using algorithms. Using Bag of Words and TF-IDF for their investigation, they observed that Bag of Words performed better than TF-IDF. In addition, they used Naive Bayes, Decision Trees, Random Forests (RFs), and Support Vector Machines and found that RF provided the most accurate categorization. Given that millions of individuals shared their thoughts about the protests, one of the study's limitations is that they may have retrieved a rather high number of tweets. A greater quantity of tweets may have been useful in revealing a variety of emotions.

Using Twitter data, Swathi *et al.* [24] provide a novel teaching and learning-based optimization (TLBO) model with long short-term memory (LSTM)-based sentiment analysis for stock price prediction. Due to the short length and peculiar grammatical patterns of tweets, data pre-processing is required to eliminate irrelevant information and put it into a readable format. In addition, the LSTM model is used to categorize tweets into positive and negative opinions about stock values. They help explore the correlation between tweets and stock market values. The Adam optimizer is used to set the learning rate of the LSTM model to enhance its prediction performance. In addition, the TLBO model is used to properly adjust the output unit of the LSTM model. On Twitter data, experiments are conducted to improve the forecasting ability of the TLBO-LSTM model for stock prices. The experimental results of the TLBO-LSTM model outperform the state-of-the-art approaches in a variety of respects. The TLBO-LSTM model gave an excellent result, with a maximum accuracy of 95.33%, a recall of 85.28%, and an F-score of 90%. The TLBO-LSTM model outperformed the competition by attaining a superior accuracy of 94.73%.

Dubey [25] used Twitter Sentiment Analysis to ascertain how residents in different countries are coping with the COVID-19 outbreak. The research analyzed tweets from 12 different countries. These tweets were gathered between March 11, and March 31, 2020, and are associated to COVID-19 in some manner. The tweets were acquired, pre-processed, and then subjected to sentiment and text mining analysis. The study's findings show that, although the most people worldwide are optimistic and hopeful, there are instances of fear, sadness, and disdain around the globe. The study analyzed tweets from the selected nations using the NRC Emtoion lexicon. The NRC Lexicon of Word-Emotion Associations has 10,170 lexical units that examine not just positive and negative polarity, but also the eight emotions established by Plutchik. On average, 50,000 tweets were used in the study from each nation every 4 days. The collection was conducted using the R package RTweet. COVID-19, coronavirus, corona, stay home stay safe, and COVID-19 pandemic were the keywords used to gather the tweets. While collecting the tweets, the retweets and responses were filtered out to prevent repetition. When the whole database was in hand, data cleaning was done,

which included the removal of white spaces, punctuation, stop words, and the conversion of tweets to lower case. Following data cleansing, the tweets were analyzed using the NRC Emtoion lexicon using the get_nrc sentiment function. After scoring tweets on feelings and emotions, a corpus was built to generate a word cloud for each nation. However, a drawback of the study is that the NRC Emtoion language does not include sarcasm and irony as emotions.

In another study, Monica and Nagarathna [26] give users who have recently written about a certain topic a model that analyzes how they feel about it based on real-time data. They use this algorithm to create a sentiment score for each user based on content-based criteria to detect Twitter spam. The suggested method applies a custom rule-based algorithm for bot detection and compares it to a number of different algorithms such as MLP, decision tree, and RF to establish the model's effectiveness in detecting spam accounts. The Twitter API was used to collect real-time data for this investigation. The data extraction procedure includes extracting the characteristics required for the research, preprocessing, and sentiment analysis. Then, using the Fake Prediction Algorithm, MLP, Decision Tree, and RF, the data are categorized to determine how many of them are authentic and legitimate users. They resulted that the rule-based fake prediction system achieved the score of accuracy of 0.97, which was superior to the existing machine learning classifiers. The study has two major limitations. First, the group of users from which data had been collected is small. Second, English was the only language examined for analysis.

Using data from the 2010 Chile earthquake and the 2017 Catalan independence vote, Ruz *et al.* [27] examined five classifiers (one of which is a variation of the TAN model) and evaluated their effectiveness on two Twitter datasets. They are considering Bayesian network classifiers for sentiment analysis on two Spanish-language datasets: The 2010 Chilean earthquake and the 2017 Catalan independence vote. To automatically manage the amount of edges supported by training instances in the Bayesian network classifier, they employ a Bayes factor technique, resulting in networks that are more realistic. Given a significant number of training instances, the findings demonstrate the efficacy of the Bayes factor measure and its competitive prediction performance when compared to support vector machines and RFs. In addition, the generated networks enable the identification of word-to-word relationships, so providing valuable qualitative information for understanding the key characteristics of event dynamics from a historical and social perspective. Even though there are not enough training examples, the research achieves that the event dynamics may be understood using qualitative information from TAN and BBF TAN. Furthermore, the generated networks may be applied to convey a tale about the important event that was studied. However, this study may be enhanced by applying the Bayesian network classifier and grounded theory.

Along the same line, Saad and Yang [28] effort to undertake a complete Twitter sentiment analysis using machine learning techniques and ordinal regression. The suggested technique comprises pre-processing tweets and then generating a relevant feature using a feature extraction method. The scoring and balancing aspects come next, and they may be categorized in a number of different ways. The suggested system uses RF, multinomial logistic regression (SoftMax), decision trees (DTs), and support vector regression (SVR) methods for sentiment analysis categorization. This system's real implementation is dependent on a Twitter dataset made available through the NLTK corpus resources. According to experimental data, the proposed solution may reliably detect ordinal regression using machine learning methods. Furthermore, the results suggest that Decision Trees outperform all other algorithms in terms of delivering the best outcomes. The proposed system consists of four key components. The first module is data acquisition, which is the method of gathering labeled tweets for sentiment analysis; the second module is preprocessing, which is the method of converting and refining tweets into a data set that might easily be used for further analysis. The third module emphases the extraction of relevant features for classification model construction. Following that, the method for balancing and evaluating tweets is presented. The final module sorts tweets into high positive, moderate positive, neutral, moderate negative, and high negative categories using a quantity of machine learning classifiers. Based on the study results, SVR and RF have almost the same accuracy, which is superior to the multinomial logistic regression classifier. The decision tree, however, is the most accurate, with a score of 91.81%. Based on the findings of the trials, the suggested model can accurately detect ordinal regression in Twitter using machine learning methods.

4.2. Fake Account Detection and Classification

Twitter bots are software-controlled automated Twitter accounts, while they are taught to perform duties similar to those carried out by regular Twitter users, such as like tweets and following other users. Twitter bots can be applied for a number of beneficial reasons, including broadcasting critical material such as weather crises in real time, publishing useful content in bulk, and producing automated direct message responses. However, Twitter bots might be used for negative purposes such spreading fake news campaigns, spamming, compromising others' privacy, and sock-puppetry. The

following paragraphs will be a survey of resent researches on Twitter bot detection and classification.

Kudugunta and Ferrara [29] used both conventional machine learning classifiers and deep learning techniques to identify bots on Twitter, both at the account and tweet levels. They used SMOTE with data augmentation using (1) Edited Nearest Neighbors (ENN) and (2) Tomek Links to address the unbalanced dataset. A collection of classifiers, including Logistic Regression, SGD Classifier, RF Classifier, AdaBoost Classifier, and MLP, was first trained using a minimum set of features. Second, they suggested a deep learning architecture, contextual LSTM, to discriminate between tweets made by actual people and those generated by bots. The design of contextual LSTM incorporates both tweet text and account metadata. It is a system with various inputs and outputs that produces accurate categorization results.

Alom *et al.* [30] also proposed two deep learning techniques using Convolutional Neural Networks (CNNs) for identifying spam on Twitter at both the account and tweet levels. First, they developed a text-based classifier composed of an Embedding and a CNN layer to determine whether or not a particular tweet belongs to a spammer. Next, they suggested a combined classifier that utilizes both a text-based classifier and a neural network on users' information for identifying spammers at the account level on Twitter. For their tests, they used two Twitter datasets and compared the performance of their proposed machine learning and deep learning-based techniques to that of current state-of-the-art machine learning and deep learning-based approaches.

Wei and Nguyen [31] used a deep learning architecture consisting of an Embedding layer, three Bidirectional LSTM layers, and a fully linked layer to produce the final output for identifying whether tweets on Twitter were created by actual individuals or bots. They attained performance comparable to that of current cutting-edge bot detection systems.

Martinelli *et al.* [32] developed a simplified deep learning method for determining if a single tweet was produced by a spammer or not. In the tests, the authors developed many MLP classifiers with a range of zero to four hidden layers. As features (inputs to MLP classifiers), word embeddings were used. After loading pre-trained word embeddings, they specifically turned each word to a numerical vector and then averaged all words in sentences-tweets.

Gong *et al.* [33] proposed a more complex deep learning architecture and feature extraction approaches for detecting fraudulent users on Dianping, a location-based social network. First, they retrieved information that may be categorized into five major groups: time-series, spatial-temporal, user-generated content, social, and demographic aspects. The time-series characteristics were then used as input for the deep learning model, which consisted of a BiLSTM layer followed by a fully connected layer with a softmax activation function. This model's output consists of two probabilities (probability of legitimate and probability of malicious). The probabilities were then employed with the other data (traditional features) to train machine learning algorithms and get the final result. Multiple machine learning techniques, including XGBoost, RF, C4.5 Decision Tree, and SVM, were taught. According to the F1-score, XGBoost produced the best classification results.

In their research, Gilani *et al.* [34] classified Twitter accounts into two categories: Automated Bots and Real Users. They gathered data using their own platform, Stweeler. They gathered 2.5–3 million tweets every day and divided their data into four subsets: 10 million, one million, one hundred thousand, and one thousand, each representing the account's popularity based on the amount of followers. For the tagging procedure, they employed human annotation and Cohen's kappa coefficient to ensure that the annotator judgments were reliable. In all, 3536 accounts were applied in the testing phase throughout the four bands. The authors retrieved 15 characteristics and used the RF classifier after completing a statistical computation. They did 5-fold cross-validation by teaching and testing in three different sets of experiments. The accuracy rate was 86.44%, the precision was 85.44%, the recall was 82.24%, and the F-measure was 83.4%. Among the 15 traits, they discovered that six rated the highest. There are two issues with this study. First, it relies on humans. Second, it did not use the content as one of the attributes while using NLP for content analyzing may enhance the accuracy level of the system.

In a further recent work, Pozzana and Ferrara [35] examined four tweet metrics to determine how bots behaved during a single activity session: the number of mentions per tweet, the distance of the text in the tweet, the percentage of retweets, and the portion of answers. This study identified behavioral distinctions between human users and bot accounts that may be utilized to enhance bot detection algorithms. For example, humans are continually visible to tweets and messages from other users when engaged in online activities, boosting their chance of engaging in social contact. The authors employed five machine learning methods (Extra Trees (ET), DT, RF, Adaptive Boosting (AB), and KNN) to assess whether tweets were created by a bot or a human. The studies used a dataset

of over 16 million tweets posted by over 2 million unique individuals. ET and RF had the greatest cross-validated average performance 86% followed by DT and AB 83% and KNN 81%. However, the research's failure to categorize whether the bot is harmful or not might be seen as a flaw.

To detect spam-bots, Wang [36] employed three graph-based and three tweet-based features. The graph-based elements (such as the user's number of friends, followers, and follower ratio) are retrieved from the user's social network, whereas the tweet-based elements (such as the number of duplicate tweets, HTTP links, and replies/mentions) are retrieved from the user's most recent 20 tweets. The dataset applied to evaluate this approach includes 25,847 persons, around 500K tweets, and approximately, 49M followers/friends taken from publicly accessible Twitter data. Several classification techniques, including Decision Tree (DT), Neural Network (NN), Support Vector Machines (SVM), Naive Bayesian (NB), and k-Nearest Neighbors, are used to detect spam bots (KNN). With 91% accuracy, 91% recall, and 91% F-measure, the NB classifier achieved the best outcomes.

Chu *et al.* [37] classified Twitter users into three groups based on attributes retrieved from tweet content, tweeting behavior, and account proprieties: bot, human, and cyborg. The authors thought that bot character is less sophisticated than human behavior. They used an entropy rate to identify the difficulty of a process, with low rates indicating a regular process, medium rates indicating a difficult process, and high rates indicating a random process. The body of the tweet is utilized to create text patterns of recognized spam on Twitter. Other account-related factors, for example the percentage of external URLs, the safety of links, the date of account registration, and so on, are also applied in the classification. The RF machine-learning algorithm is applied to assess these factors to determine whether a Twitter account is a human, bot, or cyborg. The classifier's effectiveness is tested using a dataset of 500,000 different Twitter users. The total true positive rate for this strategy was 96.0% on average.

In addition, following multiple experiments, Ferrara *et al.* [38] generated an artificial intelligence program to spot bots on Twitter depending on variations in patterns of tasks among legitimate and fake accounts. They examined two distinct data sets of Twitter users who were grouped as bots or humans manually and by a pre-existing methods. The manually validated data collection included 8.4 million tweets from 3500 human accounts and 3.4 million tweets from 5000 bots. According to the study, human users reacted to other tweets 4 to 5 times more often than bots. Over the course of an hour, genuine users become more engaged, with the proportion of responses growing. The length of human users' tweets decreased as the sessions went. According to Ferrara, the quantity of information conveyed is decreasing. The author privileges that the change is related to cognitive tiredness, which causes individuals to be less inclined to exert mental effort in developing new material over time. Bots, in contrast, exhibit no change in their engagement or the quantity of material they tweet time to time.

5. CONCLUSION

This research covers papers on the analysis of real-time Twitter data, including classification and identification of bots and real-time sentiment analysis. To do this, the literature on Twitter sentiment analysis and bot identification and classification was analyzed. In addition, the research evaluates Twitter's platform characteristics, Streaming API, and data analysis stages.

According to the publications examined for this study's sentiment analysis, several academics have used opinion analysis to determine the negative and positive feelings of Twitter users. According to the studied articles, readers' sarcasm and irony were never effectively evaluated. According to the publications examined in this article, the length of tweets and a decrease in the amount of information communicated, which may be evaluated by detecting the tweet's interactivity, are patterns of behaviors that can be used to distinguish between actual and fraudulent Twitter accounts that this paper offers researchers with information on the categories of Twitter bots. In addition, the paper analyzes current Twitter analytic techniques and latest Twitter bot detecting systems. As a follow-up to this study, our feature research will use Twitter sentiment analysis to enhance bot detection classification.

REFERENCES

[1] D. M. Kancherla. "A Hybrid Approach for Detecting Automated Spammers in Twitter". *International Educational Applied Research Journal*, vol. 3, no. 9, 2707-2719, 2019.

[2] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González and A. López-Cuevas. "A one-class classification approach for bot detection on Twitter". *Computers and Security*, vol. 91, p. 101715, 2020.

[3] O. Loyola-González, R. Monroy, J. Rodríguez, A. L. Cuevas and J. I. Sánchez. "Contrast pattern-based classification for bot detection on twitter". *IEEE Access*, vol. 7, pp. 45800-45817, 2019.

[4] N. A. Azeez, O. Atiku, S. Misra, A. Adewumi, R. Ahuja and R. Damaševičius. "Detection of malicious URLs on twitter". *Advances*

in Electrical and Computer Technologies, vol. 672, pp. 309-318, 2020.

[5] J. Chen. "*Twitter Metrics: How and Why You Should Track Them"*. Sprout Social, United States. 2021. Available from: https://sproutsocial.com/insights/twitter-metrics/. [Last accessed on Nov 2021 23].

[6] S. Arifuzzaman and N. S. Sattar. "COVID-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the USA". *Applied Sciences,* vol. 11, no. 14, p. 6128, 2021.

[7] O. Inya. "Egungun be careful, na Express you dey go: Socialising a newcomer-celebrity and co-constructing relational connection on Twitter Nigeria". *Journal of Pragmatics,* vol. 184, pp. 140-151, 2021.

[8] H. Piedrahita-Valdés, D. Piedrahita-Castillo, J. Bermejo-Higuera, P. Guillem-Saiz, J. R. Bermejo-Higuera, J. Guillem-Saiz, J. A. Sicilia-Montalvo and F. Machío-Regidor. "Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019". *Vaccines,* vol. 9, no. 1, p. 28, 2019.

[9] A. C. Breu. "From tweetstorm to tweetorials: Threaded tweets as a tool for medical education and knowledge dissemination". *Seminars in Nephrology,* vol. 40, no. 3, pp. 273-278, 2020.

[10] C. Wukich. "Connecting mayors: The content and formation of twitter information networks". *Urban Affairs Review*, vol. 58, pp. 33-67, 2020.

[11] N. Aguilar-Gallegos, L. E. Romero-García, E. G. Martínez-González, E. I. García-Sánchez and J. Aguilar-Ávilaa. "Dataset on dynamics of coronavirus on twitter". *Data in Brief*, vol. 30, p. 105684, 2020.

[12] S. Boon-Itt and Y. Skunkan. "Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study". *JMIR Public Health Surveill,* vol. 6, no. 4, p. e21978, 2020.

[13] V. Cheplygina, F. Hermans, C. Albers, N. Bielczyk and I. Smeets. "Ten simple rules for getting started on Twitter as a scientist". *PLoS Computational Biology,* vol. 16, no. 2, p. e1007513, 2020.

[14] R. Chandrasekaran, V. Mehta, T. Valkunde and E. Moustakas. "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study". *Journal of Medical Internet Research,* vol. 22, no. 10, p. e22624, 2020.

[15] P. Surowiec and C. Miles. "The populist style and public diplomacy: Kayfabe as performative agonism in Trump's Twitter posts". *Public Relations Inquiry,* vol. 10, no. 1, pp. 5-30, 2021.

[16] I. A. Mohammed and A. S. Abbas. "Twitter APIs for collecting data of influenza viruses, a systematic review". *2021 International Conference on Communication and Information Technology (ICICT)*, vol. 12, pp. 256-261, 2021.

[17] S. Wu, M. A. Rizoiu and L. Xie. "Variation across scales: Measurement fidelity under twitter data sampling". *Fourteenth International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 715-725, 2020.

[18] H. Ledford. "How Facebook, Twitter and other data troves are revolutionizing social science". *Nature*, vol. 582, no. 7812, pp. 328-330, 2020.

[19] Z. Pehlivan, J. Thièvre and T. Drugeon. "Archiving social media: The case of Twitter". *The Past Web*. Springer, Cham. pp. 43-56, 2021.

[20] I. Nazeer, S. K. Gupta, M. Rashid and A. Kumar. "Use of novel ensemble machine learning approach for social media sentiment analysis". *Analyzing Global Social Media Consumption*. Information Science Reference, Hershey. pp. 61-28, 2020.

[21] R. Al Bashaireh, M. Zohdy and V. Sabeeh. "Twitter Data Collection and Extraction: A Method and A New Dataset, the UTD-MI". *ICISDM 2020: Proceedings of the 2020 the 4th International Conference on Information System and Data Mining*, pp. 71-76, 2020.

[22] R. P. Mehta, M. A. Sanghvi, D. K. Shah and A. Singh. "Sentiment Analysis of Tweets Using Supervised Learning Algorithms". *First International Conference on Sustainable Technologies for Computational Intelligence*. vol. 1045, pp. 323-338, 2019.

[23] A. S. Neogi, K. A. Garga, R. K. Mishraa, Y. K. Dwivedib. "Sentiment analysis and classification of Indian farmers' protest using twitter data". *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100019, 2022.

[24] T. Swathi, N. Kasiviswanath and A. A. Rao. "An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis". *Applied Intelligence*, vol. 52, pp. 13675-13688, 2022.

[25] A. D. Dubey. "Twitter sentiment analysis during COVID-19 outbreak". *Jaipuria Institute of Management*, vol. 9, pp. 71-76, 2020.

[26] C. Monica and N. Nagaraju. Detection of fake tweets using sentiment analysis". *SN Computer Science,* vol. 1, no. 2, p. 89, 2020.

[27] G. A. Ruz, P. A. Henríquez and A. Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers". *Future Generation Computer Systems,* vol. 106, pp. 92-104, 2020.

[28] S. E. Saad and J. Yang. "Twitter sentiment analysis based on ordinal regression". *IEEE Access,* vol. 7, pp. 163677-163685, 2019.

[29] S. Kudugunta and E. Ferrara. "Deep neural networks for bot detection". *Information Sciences*, vol. 467, pp. 312-322, 2018.

[30] Z. Alom, B. Carminati, E. Ferrarib. "A deep learning model for Twitter spam detection". *Online Social Networks and Media,* vol. 18, p. 100079, 2020.

[31] F. Wei and U. T. Nguyen. "Twitter Bot Detection Using Bidirectional Long Short-Term Memory Neural Networks and Word Embeddings". *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA),* pp. 101-109, 2019.

[32] F. Martinelli, F. Mercaldo and A. Santone. "Social Network Polluting Contents Detection through Deep Learning Techniques". *2019 International Joint Conference on Neural Networks (IJCNN),* pp. 1-10, 2019.

[33] Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, X. Wang and X. Fu. "DeepScan: Exploiting deep learning for malicious account detection in location-based social networks". *IEEE Communications Magazine,* vol. 56, no. 11, pp. 21-27, 2018.

[34] Z. Gilani, E. Kochmar and J. Crowcroft. "Classification of twitter accounts into automated agents and human users". *Association for Computing Machinery*, vol. 17, p. 489-496, 2017.

[35] I. Pozzana, E. Ferrara. "Measuring bot and human behavioral dynamics". *Human Computer Interaction,* vol. 1, 1-11, 2018.

[36] A. H. Wang. "Detecting spam bots in online social networking sites: A machine learning approach". In: S. Foresti and S. Jajodia (Eds.), *Data and Applications Security and Privacy XXIV*. vol. 6166, pp. 335-342, 2010.

[37] Z. Chu, S. Gianvecchio, S. Jajodia H. Wang. Detecting automation of twitter accounts: Are you a human, bot, or cyborg?". *IEEE Transactions on Dependable and Secure Computing*, vol. 9, pp. 811-824, 2012.

[38] I. Pozzana and E. Ferrara. "Measuring bot and human behavioral dynamics". *Frontiers in Physics,* vol. 8, no. 125, p. 32, 2020.

p-ISSN 2521-4209
e-ISSN 2521-4217

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.6 No.(2) December 2022

2022 2722

e.mail:jst@uhd.edu.iq