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.7 No.(1) June 2023

2023 2723

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

Editorial Board

Technical

Advisory Board

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

Email: jst@uhd.edu.iq

Support Contact

UHD Technical Support

Phone: +964 773 393 5959

Email: jst@uhd.edu.iq

Contents

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Prevalence of Vitamin D Deficiency among Pregnant Women in Sulaimaneyah City-Iraq

Hasan Qader Sofihussein

Department of Pharmacy, Sulaimani Polytechnic University, Sulaimani Technical Institutes, Iraq

ABSTRACT

Hypovitaminosis D during pregnancy has a negative impact on the mother and infant's health status. The main source of Vitamin D is sunshine and ultraviolet B for most humans and food sources are often inadequate. The present work has been carried out to demonstrate the prevalence of Vitamin D deficiency among pregnant women in the Sulaimaneyah City/Kurdistan Region of Iraq. Serum samples were collected from 261 pregnant women who attended the Teaching Maternity Hospital and met inclusion criteria and were examined for 25-hydroxyvitamin D using the Roche Elecsys Vitamin D_3 assay. Different information included, including sociodemography, body mass index, and obstetric history, was collected using a specific questionnaire form. The study showed a high prevalence of hypovitaminosis D (71.3%) among pregnant women. High socioeconomic classes, blood group A^-, and advanced gestational age have been significantly associated with higher Vitamin D levels. Vitamin D deficiency is prevalent in pregnant women in Sulaimani city. Because of the many risk factors of Vitamin D deficiency and a series of health consequences, the government needs to take a step to address the problem, including raising awareness among the community about the burden of the situation and how to increase obtaining optimum Vitamin D from different sources.

Index Terms: Vitamin D, Pregnant Women, Hypovitaminosis D, Sulaimaniyah

1. INTRODUCTION

Vitamin D is one of the fat-soluble compounds that are divided into two forms ergocalciferol (D_2) and cholecalciferol (D_3) in relation to human health. Vitamin D_2 is derived from the diet, such as Cod liver oil and fatty fish while D_3 is synthesized in the skin from its precursor as exposed to ultraviolet irradiation [1]. Vitamin D in the human body is converted to 25-hydroxy Vitamin D (25(OH)D) which is a storage and circulating form of Vitamin D, and then to an active form (1,25-dihydroxy Vitamin D) by liver and kidney enzymes [2]. The classical function of Vitamin D is enhancing calcium absorption from the gut to maintain optimum calcium and phosphorus concentration in the blood, which is required to maintain many physiological functions such as muscle contraction, blood clotting, and enzyme activation [3]. Other biological activities of Vitamin D have been proposed by different studies, including enhancing insulin production, responding to many immune and inflammatory triggers, and cell growth and differentiation [4].

Over the last decades, huge numbers of articles have been published worldwide, confirming several Vitamin D health benefits [5]. The action of Vitamin D during pregnancy is still under study; however, Vitamin D is an essential element for the development of healthy fetal bone during pregnancy [5]. Vitamin D deficiency in pregnant women increases the risk of gestational diabetes mellitus and pre-eclampsia for the mother and increases the chances of being small for gestational age, neonatal rickets, and tetany

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.1-6

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: hasan.sofi@spu.edu.iq

Received: 28-09-2022 Accepted: 13-11-2022 Published: 02-01-2023

for offspring [6], [7]. Several studies reported Vitamin D deficiency in countries with plenty of sunshine for the majority of the time of the year such as India and Saudi Arabia [8], [9]. For the majority of people, getting exposure to sunshine between 09.00 AM and 03.00 PM (depending on solar time) can be considered the main source of Vitamin D [10]. Although in high altitudes because of elevation in solar angle and ambient UVB levels are mostly low, getting an optimal Vitamin D from the sunshine is unworkable, especially in the cooler season [11].

A high prevalence of Vitamin D deficiency has been reported among pregnant Chinese women [11]. Vitamin D deficiency occurs as a result of long-term inadequate intake of Vitamin D from food sources, impaired Vitamin D absorption from the intestine, liver, or kidney diseases, which affect the metabolism of Vitamin D to its active form and inadequate sun exposure. The vast majority of these cases can be corrected by determining underpinning factors associated with Vitamin D deficiency during pregnancy [12]. Studies concluded that taking Vitamin D supplementation during pregnancy must be considered to protect pregnant women and offspring from complications due to Vitamin D deficiency, [13]. In some countries, Vitamin D supplementation is offered for free for pregnant women, unfortunately, it is not available for pregnant women in Iraq.

The present study was carried out to explore the prevalence of Vitamin D deficiency among a group of pregnant women who were assumed to be a representative group of pregnant women in Sulaimani city. Moreover, the study also will try to investigate the association between Vitamin D level age, body mass index, and blood groups.

2. METHODS

2.1. Study Design and Population

The design of the present work is a cross-sectional study carried out from December 2018 to February 2019. Pre-specified inclusion criteria include pregnant women with a gestational age of more than 24 weeks and not on Vitamin D supplements even before pregnancy. Furthermore, women with a pre-pregnancy BMI of more than 35 and pregnant age more than 40-years-old were excluded from the study. The study samples were drowned by a systematic random sampling method from all patients who met inclusion criteria and visited the antenatal care unit in the Maternity Teaching Hospital in Sulaimani city. Totally, 261 pregnant women were successfully recruited to participate in the current cross-sectional study.

2.2. Data Collection

Trained persons collected data using face-to-face interviews. The questionnaire was divided into three main parts 1, sociodemographic data such as age, address, occupation, and income. 2, obstetric history, such as gravidity, and parity 3, dietary history, such as the quantity of routine milk and fish consumption recorded. Outdoor activity and exercise were considered. Sun exposure was defined as exposure to sunshine directly with uncovering body parts and not behind windows.

To control some confounding factors, which have an effect on the Vitamin D level, this study excludes pregnant women with high BMI (more than 35 kg/m^2), liver and kidney disease, and fat malabsorption disorders.

The blood sample was taken from the eligible pregnant women and centrifuged at 5000 rpm for 5 min then the serum was separated and stored at −80°C in deep freeze until they were used for analyzing serum 25 dehydroxyl Vitamin D measurement. Serum Vitamin D level was carried out using Roche Cobas e411 immunoassay analyzer using the Roche Elecsys Vitamin D$_3$ assay (Roche Diagnosis, Mannheim, Germany). A serum level of <20 ng/mL was considered Vitamin D deficiency, between 20 ng/mL and 30 ng/mL was considered insufficiency and more than 30 ng/mL was regarded to be the optimal level. Content validity was determined through a pane; experts were 12 experts; and reliability was measured using the correlation coefficient of (1 = 0.884 = 0.88.4) (statistically adequate). A pilot study was conducted with 20 pregnant women who attended Maternity Teaching Hospital.

3. RESULTS

Totally 261 pregnant women were recruited for the present study. More than 93% of the participants were at an age between 20 and 40-years-old, 3.4% were <20-year-old and the rest were above 40 years (1.3%). More than 44% of the pregnant women had a body mass index of more than 30 kg/m^2, and 32.6% of participants had normal weight Only 16.1% were categorized as obese, and 5% had morbid obesity according to the body mass index category and 1.3% of participants were underweight. The majority of the participants had an O$^+$ blood group (39.1%). In addition, 25.3% were A$^+$ and the rest had other blood groups. The majority of the pregnant women (77.8%) identified themselves as a housewife. Nearly half of the participants (46.3%) graduated from secondary school and only 29.1% of the participants had postgraduate degrees. Two hundred and eighteen (83.5) of the 261 participants were from the urban area of Sulaimani city (Table 1).

Table 1 showed the demographical data expressed as number (%), median; Chi-square was used for categorical variables and *t*-test for continuous variables. Differences were considered statistically significant at $P < 0.05$. BMI: Body mass index.

More than 70% of the cases got married at the ages of 20–29 years. The majority of the participants were in the second (55.9%) and third trimester (43.0%) of the pregnancy and only 1.1% had a gestational age of fewer than 20 weeks. A 170 (65.1%) of the 261 participants practised hijab and 34.9% had partly covering clothes. About 67.5% of the participants had more than one pregnancy and 32.5% were primigravida. The majority of the pregnant women were primipara (77.3%) and 22.7% of the participants had a history of more than one childbirth (Table 2).

Table 2 Distribution of the study sample according to reproductive history. Table 2 showed the reproductive data expressed as number (%), and median; Chi-square was used for categorical variables and a *t*-test for continuous variables. Differences were considered statistically significant at $P < 0.05$.

TABLE 1: Distribution of the study sample according to sociodemographic characteristics

Variables	Frequency	Percent	Mean±SD
AGE <20 years	9	3.4	28.8±4.96
20–29 years	127	48.6	
30–39 years	122	46.7	
40 years and more	3	1.3	
Blood group A^+	66	25.3	
B^+	50	19.2	
AB^+	21	8.0	
O^+	102	39.1	
A^-	7	2.7	
B^-	4	1.5	
AB^-	0	0.0	
O^-	11	4.2	
BMI Underweight	4	1.5	26.64±4.62
Normal	85	32.6	
Overweight	117	44.8	
Obese	42	16.1	
Morbid obese	13	5.0	
Occupation employee	58	22.2	
Non employed	203	77.8	
Educational status illiterate	6	2.3	
Read and write	12	4.6	
Primary school graduate	44	16.9	
Secondary school Graduate	121	46.3	
Postgraduate	76	29.1	
Others	2	0.8	
Residency Urban	218	83.5	
Sub urban	37	14.2	
Rural	6	2.3	

The result of the study showed a high prevalence of Vitamin D deficiency among pregnant women (71.3%). It was concluded that 18.0% were insufficient (mean = 24.46 ng/ml, S. D = 2.80) and 10.7% of the participants had sufficient serum levels of 25-dihydroxy Vitamin D (mean = 48.29 ng/ml, S. D = 20.12) (Table 3).

Table 3 showed the serum 25(OH) levels data expressed as frequency, percent (%) and mean. Vitamin D <20 ng/ml was considered deficient, between 20 ng/ml and 30 ng/ml considered insufficient and optimum levels above 30 Differences were considered statistically significant at $P < 0.05$.

According to (Table 4), the mean Vitamin D level was almost at the same level among different age groups (<20 years = 16.4, 20–29 years = 16.9, 30.39 years = 16.09), with the exception of ages more than 40 years, which was 26.4 ± 23.2. Likewise, positive blood groups had similar mean for serum Vitamin D levels ($A^+ = 19.36$, $B^+ = 15.70$, $AB^+ = 18.70$, $O^+ = 14.60$). Higher Vitamin D levels can be seen among participants with blood group A^- (mean = 33.53 ng/ml, S. D = 38.7). B^- and O^- blood groups had a mean of 8.52 ± 1.60 and 11.58 ± 8.06, respectively. A significant association was found between the blood group and Vitamin D status ($P = 0.009$). Furthermore, the result showed that higher socioeconomic status had higher Vitamin D levels with a significant association ($P = 0.007$). There were no significant differences in Vitamin D status among participants with different BMI. There were no significant differences in Vitamin D levels between pregnant women with different employment states, educational levels, and residency.

Table 4 demonstrates association between serum Vitamin D level and sociodemographic variables. Differences were considered statistically significant at $P < 0.05$.

As shown in Table 5, there was not any significant association found between serum Vitamin D levels among pregnant women of different ages at marriage. A significant association was found between gestational age and Vitamin D status (0.000), higher gestational age had higher Vitamin D levels. Pregnant women with partly covered clothes had significantly higher Vitamin D concentrations (mean = 19.04 ± 18.16, $P = 0.049$). Vitamin D levels between participants with different gravida and para did not show any significant correlation. The type of delivery has no impact on the Vitamin D level.

Table 2: Distribution of the study sample according to reproductive history.

VARIABLES	FREQUENCY	PERCENT	MEAN±SD
AGE AT MARRIEGELess than 20 years	62	23.7	22.18 ± 4.14
20- 29 years	183	70.1	
30 years and over	16	6.2	
GESTATIONAL AGELess than 20 Week	3	1.1	29.6 ± 4.39
20- 29 Week	146	55.9	
30- 39 Week	112	43.0	
DRESSINGPartly covered	91	34.9	
Fully covered	170	65.1	
GRAVIDAEqual to one	85	32.5	
More than one	176	67.5	
PARAOne and less	202	77.3	

TABLE 3: Vitamin D distribution

Vitamin d class	Frequency	Percent	Mean	S. D	95% confidence interval for mean		minimum	maximum
					Lower bound	Upper bound		
Deficient	186	71.3	9.91	4.91	9.20	10.62	0.0	19.80
Insufficient	47	18.0	24.46	2.80	23.64	25.28	20.50	29.8
Sufficient	28	10.7	48.29	20.12	40.49	56.09	30.90	98.0
Total	261	100.0		-	-	-	-	-

TABLE 4: The association of Vitamin D status with sociodemographic data

Variables	Mean±S.D	Std. error	*F*-test	*P*-value	Sig.
AGE <20 years	16.4±9.48	3.16	0.731	0.534	No significance
20–29 years	16.9±16.8	1.49			
30–39 years	16.09±11.8	1.07			
40 years and more	28.4±23.2	13.37			
Blood group A^{+}	19.36±17.7	2.19	2.910	0.009	Significance
B^{+}	15.70±12.6	1.79			
AB^{+}	18.70±12.5	2.72			
O^{+}	14.60±10.0	0.99			
A^{-}	33.53±38.7	14.66			
B^{-}	8.52±1.60	0.80			
AB^{-}	0	0			
O^{-}	11.98±8.06	2.43			
Socioeconomic status low class	13.29±10.3	1.52	5.09	0.007	Significance
Middle class	16.63±14.6	1.04			
High class	26.57±19.1	4.78			
BMI underweight	13.42±10.8	5.44	0.468	0.759	No significance
Normal	16.94±17.6	1.91			
Overweight	17.31±14.07	1.29			
Obese	15.95±10.69	1.67			
Morbid obese	12.07±6.56	1.81			
Employment employee	14.08±9.58	1.29	-1.529	0.127	No significance
Non employed	17.38±15.5	1.09			
Educational status Illiterate	17.18±18.7	7.66	0.578	0.717	No significance
Read and write	19.80±9.6	2.77			
Primary school graduate	15.53±17.1	2.58			
Secondary school graduate	16.05±14.0	1.27			
High education	18.01±14.3	1.64			
Others	5.29±0.55	0.39			
Residency Urban	17.24±14.9	1.01	1.862	0.157	No significance
Sub urban	12.56±10.8	1.78			
Rural	20.4±17.7	7.22			

TABLE 5: The association of Vitamin D status with reproductive history

Variable	Mean±S.D	Std-error	*F*-test	*P*-value	Sig.
age at marriage <20 years	18.90±19.6	2.49	0.991	0.373	No significance
20–29 years	15.89±12.7	0.93			
30 years and over	16.64±10.8	2.71			
GESTATIONAL AGE <20 week	8.05±1.85	1.07	2.063	0.000	Significance
20–29 week	17.37±15.8	1.30			
30–39 week	15.93±12.8	1.21			
Clothing partly covered	19.04±18.16	1.90	1.99	0.049	Significance
Fully covered	15.36±12.06	0.92			
Gravida equal to one	16.05±14.9	1.61	-0.460	0.646	No significance
More than one	16.94±14.4	1.08			
Para equal to one	17.40±15.4	1.08	1.550	0.122	No significance
More than one	14.07±10.6	1.39			
Type of delivery Normal vaginal delivery	15.23±10.4	1.17	0.391	0.677	No significance
Assisted delivery	15.24±14.8	3.98			
Caesarean section	16.96±13.6	1.63			

In this group of participants, Vitamin D levels significantly increased as the pregnancy progressed ($P = 0.000$). Likewise, pregnant women with partly covered clothes had a significantly higher amount of Vitamin D ($P = 0.049$) (Table 5).

4. DISCUSSION

Nowadays, Vitamin D attracts the attention of many researchers as many studies have elucidated the role of Vitamin D in various mechanisms in the body. Serum 25(OH)D level can precisely measure Vitamin D status because it is reflective of both exogenous and endogenous Vitamin D production. The work can be regarded as the first study conducted in Sulaimani City in Iraq, focusing on the prevalence of Vitamin D deficiency among pregnant women. The percentage rates of Vitamin D deficiency among pregnant women who were included in the present study were relatively very high (71.3%). Optimal Vitamin D level (25(OH)D 30 ng/ml) was observed in 10.7 percent of pregnant women. Related observations were reported in several studies carried out among South Asian pregnant women [8], [9], [11].

Exposing skin to ultraviolet B can be considered the main source of Vitamin D; therefore, the optimal level of Vitamin D among people who live in countries at or near the equator is expected which is not supported by the result of studies. Despite plenty of sunshine in the region, Vitamin D deficiency has got highly prevalent in this area. There are several factors with significant impacts on Vitamin D synthesis including geographical region, seasons, daytime, weather, air pollution, and skin pigmentation also skin covered with sunscreen [14]. A number of these factors may apply to this region. Because of the impact of cultural and religious beliefs, most of the body parts are covered with clothes, which may partially play a role in limitations of the skin exposure to sunlight that negatively can affect the optimum level of Vitamin D synthesis.

Although there are limited numbers of studies in the region, the observations were reported in Saudi Arabia, which recorded a relatively high prevalence of Vitamin D deficiency among the whole population and women including pregnant and non-pregnant ones [15], [16]. Due to the closeness of the culture and region or beliefs, these results can support our conclusion and interpretations about Vitamin D deficiency in Sulaimani City. This signifies that a tropical climate does not automatically provide optimum Vitamin D for the residents.

In this study, serum Vitamin D levels were significantly higher among pregnant women and those who do not practice hijab (covering all body parts except the face and hand).

In one study, participants were divided into three groups: 1. Receiving only dietary advice for Vitamin D from the health-care professional, 2. taking Vitamin D supplementation along with dietary advice, and 3. receiving a combination of dietary advice, supplementation and exercise in the Sports Centre. The result showed that serum Vitamin D in the first group had a negligible change with a 70% rise in the second group and in the third group vitamin level increased by 300% compared to baseline [16]. Unfortunately, outdoor exercise or activity is not common among women in the region, which may be another critical reason behind widespread Vitamin D deficiency.

A strong adverse relationship was observed between Vitamin D deficiency and obesity [17], [18]. Obesity (BMI ≥30) may increase the risk of Vitamin D deficiency because increased subcutaneous fat sequesters more Vitamin D and changes its release into the bloodstream [19]. In contrast, the relationship between body mass index and serum Vitamin D concentration did not observe this study. Higher levels of Vitamin D were seen among participants in blood group A^- and lower levels in participants in blood group B^-.

Furthemore, higher socioeconomic status had significantly higher serum levels of Vitamin D. It demonstrated that the diet of women with a low socioeconomic state is high in phytate and low in calcium leading to increase demand for Vitamin D.

The exact time of getting exposed to the sunshine to get optimum levels of Vitamin D is not provided yet because of the differences in the amount of Vitamin D, the person can get from the different latitudes, seasons, skin pigmentation, and age. However, some studies recommend that to get optimal Vitamin D through sunlight, skin (face, arms, legs, or back) should be exposed to direct sunshine twice a week for 30 min from 10:00 am to 03:00 pm [20], [21]. The dietary reference intake of Vitamin D is 400 IU for women during pregnancy.

5. CONCLUSION

Because of the combined risk factors for Vitamin D insufficiency among pregnant women in this region, the government must inform the public about the magnitude of the problem and the impact of Vitamin D deficiency on overall health. This can be accomplished by educating individuals about the benefits of receiving Vitamin D from sunlight and offering free Vitamin D supplementation through pre-conceptional counselling. Because of the high frequency of Vitamin D insufficiency in this region, as well as the huge impact of Vitamin D deficiency on health status, the findings of this study should be regarded more seriously. At present, folic acid is the sole supplement offered to pregnant women in the Sulaimani region's prenatal care facility.

REFERENCES

[1] B.W. Hollis. Circulating 25-hydroxyvitamin D levels indicative of vitamin D sufficiency: Implications for establishing a new effective dietary intake recommendation for Vitamin D. *Journal of Nutrition*, vol. 135, pp. 317-322, 2005.

[2] B. W. Hollis and C. L. Wagner. Vitamin D supplementation during pregnancy: Improvements in birth outcomes and complications through direct genomic alteration. *Molecular and Cellular Endocrinology*, vol. 453, pp. 113-130, 2017.

[3] R. Bouillon and T. Suda. Vitamin D: Calcium and bone homeostasis during evolution. *Bonekey Reports*, vol. 3, p. 480, 2014.

[4] H. Wolden-Kirk, C. Gysemans, A. Verstuyf, M. Chantal. Extraskeletal effects of Vitamin D. *Endocrinology and Metabolism Clinics of North America*, vol. 41, no. (3), pp. 571-594, 2012.

[5] L. S. Weinert and S. P. Silveiro. Maternal-fetal impact of Vitamin D deficiency: A critical review. *Maternal and Child Health Journal*, vol. 19, no. 1, pp. 94-101, 2014.

[6] N. Principi, S. Bianchini, E. Baggi and S. Esposito. Implications of maternal Vitamin D deficiency for the fetus, the neonate and the young infant. *European Journal of Nutrition*, vol. 52, no. 3, pp. 859-867, 2013.

[7] E. E. Delvin, B. L. Salle, F. H. Glorieux, P. Adeleine and L. S. David. Vitamin D supplementation during pregnancy: Effect on neonatal calcium homeostasis. *The Journal of Pediatrics*, vol. 109, pp. 328-334, 1986.

[8] N. A. Al-Faris. High prevalence of Vitamin D deficiency among pregnant Saudi women. *Nutrients*, vol. 8, no. 2, 6-15, 2016.

[9] H. J. W. Farrant, G. V. Krishnaveni, J. C. Hill, B. J. Boucher, D. J. Fisher, K. Noonan and C. Osmond. Vitamin D insufficiency is common in Indian mothers but is not associated with gestational diabetes or variation in newborn size. *European Journal of Clinical Nutrition*, vol. 63, no. 5, pp. 646-652, 2009.

[10] A. R. Webb and O. Engelsen. Calculated ultraviolet exposure levels for a healthy Vitamin D status. *Photochemistry and Photobiology*, vol. 82, pp. 1697-1703, 2006.

[11] C. Yun, J. Chen, Y. He, D. Mao, R. Wang, Y. Zhang and X. Yang. Vitamin D deficiency prevalence and risk factors among pregnant Chinese women. *Public Health Nutrition*, vol. 20, no. 10, pp. 1746-1754, 2017.

[12] A. Dawodu and H. Akinbi. Vitamin D nutrition in pregnancy: Current opinion. *International Journal of Womens Health*, vol. 5, pp. 333-343, 2013.

[13] C. Palacios, L. M. De-Regil, L. K. Lombardo and J. P. Peña-Rosas. Vitamin D supplementation during pregnancy: An updated meta-analysis on maternal outcomes. *Journal of Steroid Biochemistry and Molecular Biology*, vol. 164, pp. 148-155, 2016.

[14] M. F. Holick and T. C. Chen. Vitamin D deficiency: A worldwide problem with health consequences. *The American Journal of Clinical Nutrition*, vol. 87, no. 4, pp. 1080-1086, 2008.

[15] M. Al-Zoughool, A. AlShehri, A. Alqarni, A. Alarfaj and W. Tamimi. Vitamin D status of patients visiting health care centers in the coastal and Inland Cities of Saudi Arabia. *Journal of Public Health and Development Series*, vol. 1, pp. 14-21, 2015.

[16] M. Tuffaha, C. El Bcheraoui, F. Daoud, H. A. Al Hussaini, F. Alamri, M. Al Saeedi, M. Basulaiman, Z. A. Memish, M. A. AlMazroa, A. A. Al Rabeeah and A. H. Mokdad. Deficiencies under plenty of suns: Vitamin D status among adults in the kingdom of Saudi Arabia, 2013. *North American Journal of Medical Sciences*, vol. 7, pp. 467-475, 2015.

[17] H. Alfawaz, H. Tamim, S. Alharbi, S. Aljaser and W. Tamimi. Vitamin D status among patients visiting a tertiary care centre in Riyadh, Saudi Arabia: A retrospective review of 3475 cases. *BMC Public Health*, vol. 14, p. 159, 2014.

[18] A. H. Al-Elq, M. Sadat-Ali, H. A. Al-Turki, F. A. Al-Mulheim and A. K. Al-Ali. Is there a relationship between body mass index and serum Vitamin D levels? *Saudi Medical Journal*, vol. 30, pp. 1542-1546, 2009.

[19] S. Konradsen, H. Ag, F. Lindberg, S. Hexeberg and R. Jorde. Serum 1,25-dihydroxy Vitamin D is inversely associated with body mass index. *European Journal of Nutrition*, vol. 47, pp. 87-91, 2008.

[20] M. F. Holick, T. C. Chen, Z. Lu and E. Sauter. Vitamin D and skin physiology: A D-lightful story. *Journal of Bone and Mineral Research*, vol. 22, pp. V28-33, 2007.

[21] M. F. Holick. Sunlight and Vitamin D for bone health and prevention of autoimmune diseases, cancers and cardiovascular disease. *The American Journal of Clinical Nutrition*, vol. 80, pp. 1678S-1688S, 2004.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Computer-aided Diagnosis for the Early Breast Cancer Detection

Miran Hakim Aziz[1], **Alan Anwer Abdulla**[2,3]

[1]Applied Computer, Collage of Medicals and Applied Sciences, Charmo University, Chamchamal, 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 development of the use of medical image processing in the healthcare sector has contributed to enhancing the quality/accuracy of disease diagnosis or early detection because diagnosing a disease or cancer and identifying treatments manually is costly, time-consuming, and requires professional staff. Computer-aided diagnosis (CAD) system is a prominent tool for the detection of different forms of diseases, especially cancers, based on medical imaging. Digital image processing is a critical in the processing and analysis of medical images for the disease diagnosis and detection. This study introduces a CAD system for detecting breast cancer. Once the breast region is segmented from the mammograms image, certain texture and statistical features are extracted. Gray level run length matrix feature extraction technique is implemented to extracted texture features. On the other hand, statistical features such as skewness, mean, entropy, and standard deviation are extracted. Consequently, on the basis of the extracted features, support vector machine and K-nearest neighbor classifier techniques are utilized to classify the segmented region as normal or abnormal. The performance of the proposed approach has been investigated through extensive experiments conducted on the well-known Mammographic Image Analysis Society dataset of mammography images. The experimental findings show that the suggested approach outperforms other existing approaches, with an accuracy rate of 99.7%.

Index Terms: Computer-aided Diagnosis, Medical Image, Breast Cancer, Gray Level Run Length Matrix, Classifier Technique

1. INTRODUCTION

Digital image processing (DIP) is significant in many areas, particularly medical image processing, image in-painting, pattern recognition, biometrics, content-based image retrieval, image de-hazing, and multimedia security [1], [2]. It is becoming more important for analyzing medical images and identifying abnormalities in these images. Computer-aided diagnosis (CAD) systems based on image processing have emerged as an intriguing topic in the field of medical image processing research. A CAD system is a computer-based system that assists medical professionals in diagnosing diseases, in particular cancers, using medical images such as X-ray, magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and microscopic images [3]. The aim of developing autonomous CAD systems is to extract the targeted illnesses with a high accuracy and at a lower cost and time consumption. Preprocessing, segmentation, feature extraction, and classification are the four basic phases of each CAD system. A feature is an important factor to categorize the disease in the cancer detection systems. Feature extraction is the process of transforming raw data into a set of features [4]. There are numerous types of cancers such as breast cancer, brain tumors, lung cancer, skin cancer, and blood cancer. This paper focuses on the early detection of the cancerous cells in the breast. Breast cancer is one of the most frequent kinds of cancer among females worldwide. There are currently no strategies for preventing breast cancer. The difficulty of radiologist interpretation of mammogram images can

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp7-14	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Dr. Alan Anwer Abdulla, Assistant Prof., Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq, Department of Information Technology, University College of Goizha, Sulaimani, Iraq. E-mail: Alan.abdulla@univsul.edu.iq

Received: 17-09-2022 Accepted: 11-12-2022 Published: 12-01-2023

be alleviated by employing the early-stage breast cancer detection method. Thus, early diagnosis of this condition is critical in its treatment and has a significant influence in minimizing mortality. The most effective way of detecting breast cancer in its early stages is to analyze mammography images [5]. Breast cancer is a disorder in which the cells of the breast proliferate uncontrollably. The kind of breast cancer is determined by which cells in the breast develop into cancer. Breast cancer can start in any part of the breast. It can spread outside of the breast through blood and lymph arteries. Breast cancer is considered to have metastasized when it spreads to other regions of the body [6]. In general, a breast is composed of three major components: lobules, ducts, and connective tissue (Fig. 1) [6].

The lobules are the milk-producing glands. Ducts are tubes that transport milk to the nipple. The majority of breast cancers start in the lobules or ducts [6]. Connective tissue joins or separates and supports all other forms of bodily tissue. It contains of cells surrounded by a fluid compartment termed the extracellular matrix (ECM), as do all other forms of tissue. However, connective tissue varies from other kinds in that its cells are loosely instead of densely packed inside the ECM [7].

The aim of this study is developing a CAD system for the early detection of breast cancer. The developed CAD system has the advantages of increasing accuracy rate, reducing time consumption, and reducing cost in comparison with manually detecting system. The main contributions of the proposed approach are segmenting the breast region properly as well as extracting the most significant features, and this leads to increase the accuracy rate and reduce mistake rate of wrongly treating patients. The proposed system includes the following steps: A pre-processing step for enhancing the image quality, a segmentation step for segmenting the breast region from the other components of mammography images, and a feature extraction step for extracting the most influential features. Finally, the classification step is conducted, which helps the system decide whether a cell is cancerous or non- cancerous. The rest of the paper is structured as follows. Section 2 provides a summary of past efforts from the literature. Section 3 presents the proposed CAD system. Section 4 shows the results of experiments. Finally, Section 5 gives the conclusion.

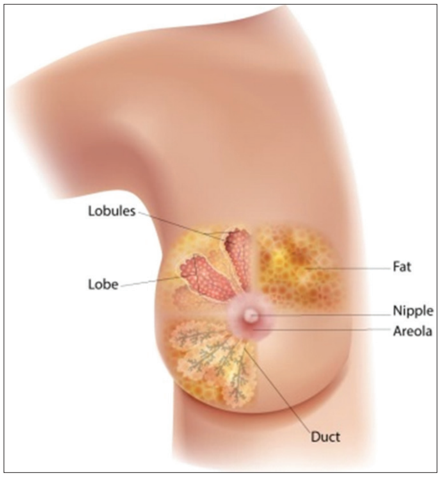


Fig. 1. Major components of the breast [6].

2. LITERATURE REVIEW

In medical image processing, the CAD system is a computer-based system that helps clinicians in their last decision about different diseases, especially cancers. The whole process is about extracting significant information from medical images such as: MRI, CT, and ultrasounds. Several CAD systems have been developed for identifying different diseases including: Breast cancer, tumor detection, and lung cancer. This study concentrates on breast cancer.

The processing and analysis of breast mammogram images plays a significant role in the early diagnosis of breast cancer. This section reviews the most influential as well as relevant current efforts on the early breast cancer detection using DIP. The main obstacle in this field of research is reducing the rate of breast cancer detection errors. In general, most of the CAD systems for the early breast cancer detection consist of the following steps: Image enhancement, image segmentation, feature extraction, feature selection, and classification.

In 2010, Eltoukhy *et al.* suggested an algorithm for the breast cancer detection using a curvelet transform technique at multiple scales [8]. Different scales of the largest curvelet coefficients are extracted and investigated from each level as a classification feature vector. This algorithm is reached an accuracy rate of 98.59% at Scale 2. Srivastava *et al.*, in 2013, introduced a CAD system for the early breast cancer diagnosis using digital mammographic images [9]. Contrast-limited histogram equalization technique is utilized for the enhncement purposes. Consequently, three-class fuzzy C-means is used for the segmentation process. The texture features such as geometric/shape, wavelet-based, and Gabor were extracted. The minimum redundancy maximum relevance feature selection method was utilized to select the fewest redundant and most relevant characteristics. Finally, Support Vector

Machine (SVM), K-Nearest Neighbor (kNN), and Artificial Neural Network (ANN) classifier techniques were used for classifying cancerious and non-canceroius cells. Furthermore, SVM provides better results in comparison to the kNN and ANN. This technique is achieved an accuracy rate of 85.57% for the 10-fold cross-validation using Mammographic Image Analysis Society (MIAS) dataset of images.

Vishrutha *et al*., in 2015, developed a strategy for combining wavelet and texture information that leads to increase the accuracy rate of the developed CAD system for the early breast cancer diagnosis [10]. The mammogram images were pre-processed using median filter. In addition, the label and the black background are removed on the bases of sum of each column's intensities. Consequently, if the total intensity of a column falls below a certain level/threshold, the column will be removed. The resulted images from the pre-processing step were utilized as input for the region growth technique used to determine the region of interest (ROI) as a seqmentation step. Discrete Wavelet Transform technique was used to extract features from the seqmented images/regions. Finally, SVM classifier technique was utilized to categorize the mammogram images as benign or malignant with an accuracy rate of 92% using Mini-MIAS dataset of images.

In 2017, Pashoutan *et al*. developed a CAD system for the early breast cancer diagnosis [11]. For the pre-processing step, cropping begins by employing coordinates and an estimated radius of any artifacts introduced into images to get to the ROI where bulk and aberrant tissues are found. Moreover, histogram equalization and median filter were used to enhance the contrast of the images. Edge-based segmentation and region-based segmentation methods are that the two main methods were used for the segmentation purposes. Furthermore, four different techniques were utilized for extracting features, such as Wavelet transform, Gabor wavlet transform, Zernike moments, and Gray-Level Cooccurance Matrix (GLCM). Eventually, using the MIAS dataset, this technique reached an accuracy rate of 94.18%.

Hariraj *et al*., in 2018, developed a CAD system for the breast cancer detection [12]. In the pre-processing step, Fuzzy Multi-layer was used to eliminate background information such as labels and wedges from images. Moreover, thresholding was used to transform the grayscale image to the binary image. Furthermore, morphological technique was implemented on the binary image to remove undesirable tiny items. Regarding to the segmentation step, K-means clustering was utilized. For the feature extraction purposes, certain shape and texture features were extracted such as: diameter, perimeter, compactness, mean, standard deviation, entropy, and correlation. Finally, the Fuzzy Multi-Layer SVM classifier technique provides better accuracy rate of 98% out of other tested classifier techniques using Mini-Mammographic MIAS dataset of images.

Sarosa *et al*., in 2019, designed a breast cancer diagnosis technique by investigating GLCM and Backpropagation Neural Network (BPNN) classification technique [13]. Histogram equalization was utilized for the pre-processing and enhancing the images. Consequently, GLCM was used to extract features from the pre-processed images. Finally BPNN was used to determine whether the input image is normal or abnormal. The suggested approach was evaluated using a MIAS dataset of images and it achieved an accuracy rate of 90%.

In 2019, Arafa *et al*. introduced a technique for the breast cancer detection [14]. In the pre-processing step, just the area including the breast region is automatically picked and artifacts as well as pectoral muscle were removed. The Gaussian Mixture Model (GMM) was utilized to extract the ROI. Moreover, texture, shape, and statistical features were extracted from the ROI. For the texture feature, GLCM was utilized. Furthermore, the following shape features such as circularity, brightness, compactness, and volume were extracted. Regarding to the statistical features, mean, standard deviation, correlation, skewness, smoothness, kurtosis, energy, and histogram were extracted. Finally SVM classifier technique was used to classify segmented ROI into normal, abnormal, benign, and malignant. This proposed technique was evaluated using MIAS dataset of images and it achieves an accuracy of 92.5%.

Farhan and Kamil developed a CAD system for classifying the input mamogram images into normal or abnormal, in 2020, [15]. At the beginning, contrast limited adaptive histogram equalization (CLAHE) method was used to improve all mammogram images. In addition, the histogram of oriented gradient, GLCM, as well as the local binary pattern (LBP) techniques was used to extract features. Finally, SVM and kNN classifier techniques were used for classifying cancerious and non-canceroius cells. The best accuracy rate of 90.3%, using Mini-MIAS dataset, was obtained when GLCM and kNN were used.

In 2020, Eltrass and Salama developed a technique for breast cancer diagnosis [16]. As a pre-processing step, the mammography image was translated into a binary image, and then all regions are sorted to identify the mammogram's greatest area, that is, breast region. In addition, all artifacts and pectoral muscle were eliminated. This CAD system utilized the expectation maximization technique for the segmentation purposes. Wavelet-based contourlet transform technique was used to extract features. Finally, SVM classifier technique was used and an accuracy rate of 98.16% was achieved using MIAS dataset.

Saeed *et al.*, in 2020, designed a classifier model to aid radiologists in providing a second opinion when diagnosing mammograms [17]. In the pre-processing step, median filter was used to remove noise and minor artifacts. Hybrid Bounding Box and Region Growing algorithm was used to segment the ROI. For the features extraction, two types of features were extracted which are: (1) Statistical features such as mean, standard deviation skewness, and kurtosis and (2) texture features such as LBP and GLCM. Consequently, SVM was used to categorize mammography images as normal or abnormal in the first level, and benign or malignant in the second level. This proposed technique used MAIS dataset to evaluate the performance, and an accuracy of 95.45% was obtained for the first level and 97.26% for the second level.

Mu'jizah and Novitasari in 2021, developed a CAD system for the breast cancer diagnosis [18]. At the beginning, certain pre-processing techniques, such as Gaussian filter and Canny edge detection technique, were implemented to enhance the visual quality of the input images. The thresholding method was also used for the segmentation purposes. To extract features, GLCM was used as texture feature, and area, perimeter, metric, as well as eccentricity were extracted as shape feature. Finally, for the classification step, SVM was used and an accuracy rate of 98.44% was obtained using Mini-MIAS dataset of images.

Recently, in 2022, Holi produced a breast cancer detection system [19] which used a median filter and CLAHE for enhancing the input image. Then, Chebyshev Distanced-Fuzzy C-Means Clustering was used to segment the pre-processed image. The augmented local vector pattern, shape features, and GLCM were used to extract features. The classification step was conducted using kNN classifier technique. This proposed technique was achieved an accuracy rate of 97% using MIAS dataset of images.

The remainder of this paper concerns with the extension and further refinement of the strategy of using DIP to increase the accuracy rate for the early breast cancer detection.

3. PROPOSED APPROACH

The microscopic image of breast is called a mammogram, which consists of three parts/regions. The breast part appears on a mammogram in colors of gray and white, while the mammogram backdrop is often black. In addition, a lump or tumor appears as a concentrated white area. Tumors may be either malignant or benign [20]. The most significant step of each CAD system for the breast cancer detection is extracting/cropping the ROI from the other parts of the mammogram image. This section describes the proposed approach which involves the following steps:

1. Pre-processing: In this step, certain techniques are applied such as region-props to delete the label from the mammogram images, and median filter as well as adaptive histogram equalization to enhance the image quality (Fig. 2).
2. Segmentation: To segment the ROI from other parts of the input image, the thresholding segmentation technique is applied on image (d) in Fig. 2, and the resulted image is a binary image, see image (a) in Fig. 3.

The threshold-based segmentation approach is an effective segmentation technique that divides an image based on the intensity value of each pixel. It is used to segment an image into smaller portions using a single color value to generate a binary image, with black representing the background and

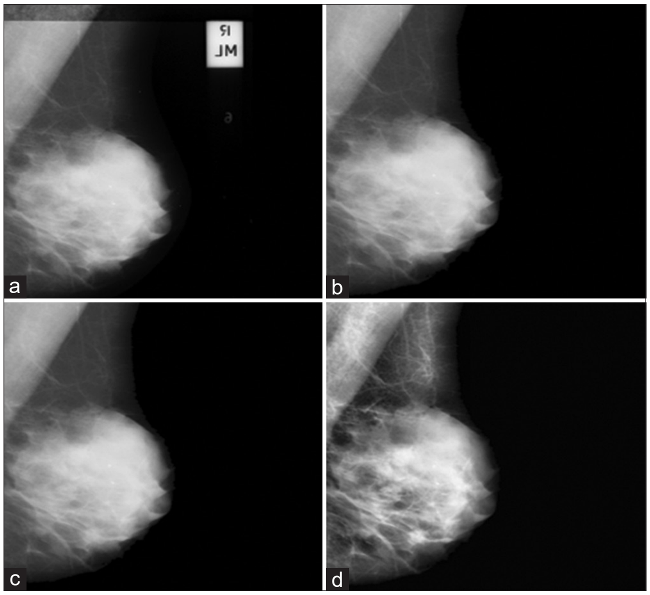


Fig. 2. Pre-Processing Step:(a) Original Mammogram Image, (b) Label Removed, (c) Resulted Image After the Median Filter has been applied on Image (b), and (d) Resulted Image after Histogram Equalization has been applied on Image (c).

white representing the objects [21]. The threshold *T* value can be selected either manually or automatically based on the characteristics of the image. In the proposed approach, $T = 0.7$ was used, which provides the optimum accuracy results. In the next section, all the tested values for the *T* are illustrated in Table 5.

3. Feature Extraction: Texture features and statistical features are extracted from the segmented image, that is, image (b) in Fig. 3. The extracted features are summarized in Table 5. Furthermore, all the extracted features are fused for the classification purposes.
4. Classification: SVM and kNN classification techniques were applied on the extracted features to distinguish normal cells from abnormal cells. The reason behind using SVM and kNN is because these two classifier techniques are the most common used in this field of research. For the both classifiers, the k-fold cross-validation with k = 5, 10, 15, and 20 was investigated.

Fig. 4 illustrates the block diagram of the proposed approach.

4. EXPERIMENTAL RESULTS

The primary goal of the proposed CAD approach is classifying the breast cancer cells into normal or abnormal. Experiments are carried out in a thorough manner in this part of the study to evaluate how well the suggested approach works in terms of accuracy rate. In addition, the proposed approach is assessed alongside the findings of the earlier research.

4.1. Dataset

The MIAS dataset provides the tested input images, which are taken from the public domain and are quite well recognized. The MIAS dataset contains the original 322 images, 206 normal and 116 abnormal, in the PGM format [22]. All of the images have the same resolution which is 1024 by 1024 pixels. The MIAS dataset has been taken into consideration in order to assess the performance of the proposed CAD approach.

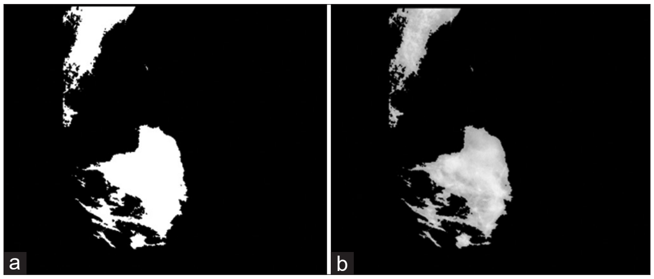


Fig 3. Segmentation step: (a) Binary image, (b) Based on the binary image in (a), the ROI is selected in the original image.

4.2. Results

Using several classifier techniques, such as SVM and kNN, the accuracy rate for each the extracted features is assessed. Tables 2 and 3 present the accuracy rate of statistical and GLCM separately using SVM and kNN respectively. In all the evaluation tests, different values ok *k-fold* have been considered. In addition, the accuracy rate has been calculated using the following formula [23]:

$$\text{Accuracy rate} = TP + TN/(TP + TN + FP + FN) \quad (1)$$

Where: TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative, respectively.

More investigation has been conducted by fusing the extracted features, namely statistical and GLCM. Meanwhile, the 11 retrieved features are utilized to evaluate the effectiveness

TABLE 1: Extracted features

Type of Features	Name of the Feature
Statistical Features	Skewness Mean Entropy Standard deviation
Texture feature: Gray level run length matrix	Short run emphasis Long run emphasis Gray level non-uniformity Run percentage Run length non-uniformity Low gray level run emphasis High gray level run emphasis

TABLE 2: SVM-based accuracy rate for the extracted features separately

Features	5 K-fold	10 K-fold	15 K-fold	20 K-fold	Average (%)
Statistical	99.1	99.2	98.8	99.3	99.1
GLRLM	98.1	99.3	99.1	99.1	98.1

SVM: Support vector machine, GLRLM: Gray level run length matrix

TABLE 3: kNN-based accuracy rate for the extracted features separately

Features	5 K-fold	10 K-fold	15 K-fold	20 K-fold	Average (%)
Statistical	94.4	94.7	97.5	97.5	96
GLRLM	97.2	98.1	97.6	98.3	97.8

kNN: K-nearest neighbor, GLRLM: Gray level run length matrix

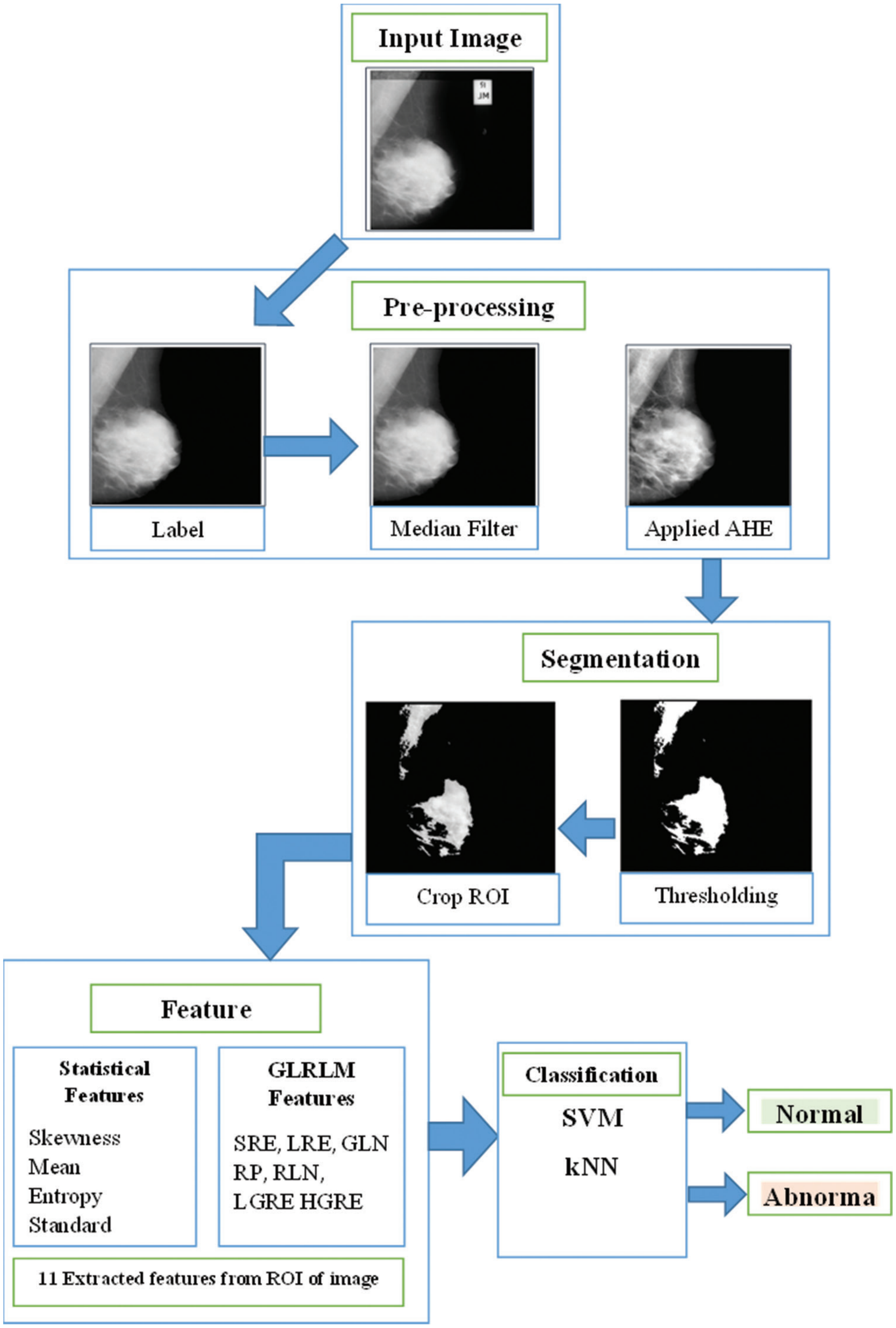


Fig. 4. Block diagram of the proposed approach.

of the proposed CAD approach in distinguishing between normal and abnormal cells. Those 11 features are previously mentioned in Table 1. Moreover, *k*- fold cross-validation with various values of *k* is used in the evaluation process to measure the accuracy. Training and testing have been done using *k*-fold cross-validation, which divides data automatically into training and testing depending on the value of *k*.Based on the investigation conducted in this study, the SVM classifier technique provides a higher accuracy rate (Table 4).

Tables 5 and 6 illustrate the findings of further tests done by comparing the obtained results of the proposed approach to results of four existing approaches. Two of the existing works were used SVM classifier techniques and the remained two works were used kNN. All of the four tested CAD systems used only 5k fold to evaluate the performance of their approaches and also Tables 7 illustrate the Time consumption of the all process in our system.

TABLE 4: Accuracy rate of the proposed CAD approach

	Cross validation				
	5K	10K	15k	20k	Average (%)
SVM	99.7	99.8	99.4	99.7	99.7
kNN	98.4	99.1	98.8	98.8	98.9

CAD: Computer-aided diagnosis, SVM: Support vector machine, kNN: K-nearest neighbor

TABLE 5: Accuracy rate of the tested approaches using SVM

	Accuracy Rate (%)
Proposed	99.7
Mu'jizah and Novitasari[18]	98.4
Eltrass and Salama [16]	98.1

TABLE 6: Accuracy rate of the tested approaches using K-nearest neighbor

	Accuracy Rate (%)
Proposed	98.9
Farhan and Kamil[15]	90.3
Holi [19]	97

Table 7: Time consumption of the proposed computer-aided diagnosis system

Stage	Times in second
Pre-processing	0.371
Segmentation	0.298
Feature extraction	0.061
Classification	1.274
Total	2.004 s

TABLE 8: Investigating the optimum value for thresholding *T*

Thresholding values	kNN (%)	SVM (%)
0.1	89	90.8
0.2	89.7	91.1
0.3	89.8	91.9
0.4	94.1	94.7
0.5	96.3	96
0.6	97.6	99.1
0.7	98.9	99.7
0.8	98.4	99.3
0.9	98.6	99.5

kNN: K-nearest neighbor, SVM: Support vector machine

According to the results presented in Tables 5 and 6, the best accuracy rate is achieved by the proposed approach and it outperforms all the tested existing approaches. Moreover, in Eltrass and Salama [16], the total time consumption is highlighted which is (2.26267) second, while the time consuming of our proposed approach is (2.004) second. The time consumption of the proposed approach is calculated as follows:

More investigations have been done for testing the optimum value for the thresholding *T* that used for the segmentation purposes. Based on the results presented in Table 8, it is quite obvious that the best accuracy rate was achieved when $T = 7$.

5. CONCLUSIONS

Since detecting a disease/cancer and identifying treatments manually is costly, time consuming, and requires professional staff, the evolution of the application of medical image processing in the healthcare field has contributed in an improvement in the quality/accuracy of disease diagnosis (or early detection). Meanwhile, medical image processing techniques can accurately extract target diseases/cancers at higher accuracy and lower cost. Breast cancer is one of the leading causes of mortality among women, compared to all other cancers. Therefore, early detection of breast cancer is necessary to reduce fatalities. Thus, early detection of breast cancer cells may be anticipated using recent machine learning approaches. The primary objective of developing CAD system for mammogram images is to aid physicians and diagnostic experts by providing a second perspective, this increases confidence in the diagnostic process. This study was focused on the development of an efficient CAD system for the early breast cancer detection. The testing findings reveal that the proposed CAD approach obtained an accuracy rate of 99.7% and outperforms the existing approaches.

To improve the performance of the proposed approach, the following are points of potential plans that extend our work in the future: (1) More filters and image processing techniques will be tested for pre-processing purposes to enhance the image quality, (2) different techniques will be tested to improve segmenting purposes, and (3) different kinds of features should be tested and investigated.

REFERENCES

[1] A. A. Abdulla. "Efficient computer-aided diagnosis technique for leukaemia cancer detection". *The Institution of Engineering and Technology*, vol. 14, no. 17, pp. 4435-4440, 2020.

[2] A. A. Abdulla and M. W. Ahmed. "An improved image quality algorithm for exemplar-based image inpainting". *Multimedia Tools and Applications*, vol. 80, pp. 13143-13156, 2021.

[3] H. Arimura, T. Magome, Y. Yamashita and D. Yamamoto. "Computer-aided diagnosis systems for brain diseases in magnetic resonance images". *Algorithms*, vol. 2, no. 3, pp. 925-952, 2009.

[4] G. Kumar and P. K. Bhatia. "A Detailed Review of Feature Extraction in Image Processing Systems". *International Conference on Advanced Computing and Communication Technologies ACCT*, pp. 5-12, 2014.

[5] T. T. Htay and S. S. Maung. "Early Stage Breast Cancer Detection System Using GLCM Feature Extraction and K-Nearest Neighbor (k-NN) on Mammography Image". *2018-The 18th International Symposium on Communications and Information Technologies*, pp. 345-348, 2018.

[6] Centers for Disease Control and Prevention. "*What Is Breast Cancer*?". Centers for Disease Control and Prevention, United States. 2021. Available from: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.html [Last accessed on 2022 Dec 18].

[7] J. Vasković. "*Overview and Types of Connective Tissue*." Medical and Anatomy Experts, 2022. Available from: https://www.kenhub.com/en/library/anatomy/overview-and-types-of-connective-tissue [Last accessed on 2022 Dec 20].

[8] M. M. Eltoukhy, I. Faye and B. B. Samir. "Breast cancer diagnosis in digital mammogram using multiscale curvelet transform". *Computerized Medical Imaging and Graphics*, vol. 34, no. 4, pp. 269-276, 2010.

[9] S. Srivastava, N. Sharma, S. K. Singh and R. Srivastava. "Design, analysis and classifier evaluation for a CAD tool for breast cancer detection from digital mammograms". *International Journal of Biomedical Engineering and Technology*, vol. 13, no. 3, pp. 270-300, 2013.

[10] S. C. Satapathy, B. N. Biswal, S. K. Udgata and J. K. Mandal. "Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014". *Advances in Intelligent Systems and Computing*, vol. 327, pp. 413-419, 2014.

[11] S. Pashoutan, S. B. Shokouhi and M. Pashoutan. "Automatic Breast Tumor Classification Using a Level Set Method and Feature Extraction in Mammography." *2017 24th Iranian Conference on Biomedical Engineering and 2017 2nd International Iranian Conference on Biomedical Engineering ICBME 2017*, pp. 1-6, 2018.

[12] V. Hariraj, W. Khairunizam, V. Vijean and Z. Ibrahim. "Fuzzy multi-layer SVM classification". *International Journal of Mechanical Engineering and Technology (IJMET)*, vol. 9, pp. 1281-1299, 2018.

[13] S. J. A. Sarosa, F. Utaminingrum and F. A. Bachtiar. "Breast cancer classification using GLCM and BPNN". *International Journal of Advances in Soft Computing and Its Applications*, vol. 11, no. 3, pp. 157-172, 2019.

[14] A. Arafa, N. El-Sokary, A. Asad and H. Hefny. "Computer-aided detection system for breast cancer based on GMM and SVM". *Arab Journal of Nuclear Sciences and Applications*, vol. 52, no. 2, pp. 142-150, 2019.

[15] A. H. Farhan and M. Y. Kamil. "Texture analysis of breast cancer via LBP, HOG, and GLCM techniques". *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 7, p. 072098, 2020.

[16] A. S. Eltrass and M. S. Salama. "Fully automated scheme for computer-aided detection and breast cancer diagnosis using digitised mammograms". *IET The Institution of Engineering and Technology*, vol. 14, no. 3, pp. 495-505, 2020.

[17] E. M. H. Saeed, H. A. Saleh and E. A. Khalel. "Classification of mammograms based on features extraction techniques using support vector machine". *Computer Science and Information Technologies*, vol. 2, no. 3, pp. 121-131, 2020.

[18] H. Mu'jizah and D. C. R. Novitasari. "Comparison of the histogram of oriented gradient, GLCM, and shape feature extraction methods for breast cancer classification using SVM". *Journal of Technology and Computer Systems*, vol. 9, no. 3, pp. 150-156, 2021.

[19] G. Holi. "Automatic breast cancer detection with optimized ensemble of classifiers". *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 11, pp. 2545-2555, 2020.

[20] V. R. Nwadike. "*What Does Breast Cancer Look Like on a Mammogram*?". 2018. Available from: https://www.medicalnewstoday.com/articles/322068 [Last accessed on 2022 Dec 16].

[21] K. Bhargavi and S. Jyothi. "A survey on threshold based segmentation technique in image processing". *International Journal of Innovative Research and Development*, vol. 3, no. 12, pp. 234-239, 2014.

[22] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, N, S. Kok, P. Taylor, D. Betal and J. Savage. "The mammographic image analysis society digital mammogram database". *International Congress Series*, vol. 1069, pp. 375-378, 1994.

[23] R. Murtirawat, S. Panchal, V. K. Singh and Y. Panchal. "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning". *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, pp. 534-540, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Transformer-based Neural Network Machine Translation Model for the Kurdish Sorani Dialect

Soran Badawi

Language Center, Charmo Center for Scientific Research & Consulting, Charmo University, Chamchamal, Sulaimani, KRG, Iraq

ABSTRACT

The transformer model is one of the most recently developed models for translating texts into another language. The model uses the principle of attention mechanism, surpassing previous models, such as sequence-to-sequence, in terms of performance. It performed well with highly resourced English, French, and German languages. Using the model architecture, we investigate training the modified version of the model in a low-resourced language such as the Kurdish language. This paper presents the first-ever transformer-based neural machine translation model for the Kurdish language by utilizing vocabulary dictionary units that share vocabulary across the dataset. For this purpose, we combine all the existing parallel corpora of Kurdish – English by building a large corpus and training it on the proposed transformer model. The outcome indicated that the suggested transformer model works well with Kurdish texts by scoring (0.45) on bilingual evaluation understudy (BLEU). According to the BLEU standard, the score indicates a high-quality translation.

Index Terms: Machine translation, Transformers, Dialect, Kurdish language, Bilingual evaluation understudy

1. INTRODUCTION

Human language has a complex and irregular system that can pose significant issues for machine translation. The kind of morphemes, their implications, and their syntactic and semantic relations in the context is causing the natural language to be complex and abnormal. The complexity of these problems has led some to believe that human translation is infeasible for such tasks.

Historically, machine translation has experienced numerous changes. First, dictionary-based and rule-based translation methods were developed and provided translation services through the manual specification of rules and resources [1]. Following that, statistical translation emerged as the new model to diminish the role of a linguist and increase the emphasis on language dependency [2].

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp15-21	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Luckily, the advancement of neural networks and artificial intelligence has primarily impacted many areas of science, including machine translation. As a result of the neural machine translation research, top-notch translations were produced for texts written in resourceful languages. Therefore, the need to achieve the same goal for low-resourced languages has become significant and the attempts to achieve that have increased [3]. Languages are considered less resourced when they lack human-constructed linguistic resources, substantial monolingual or parallel corpora, and general-purpose grammar are the sole sources available. The research industry has primarily ignored Kurdish dialects, which are practiced by 20–30 million people across four regions [4].

Corresponding author's e-mail: Soran Badawi, Language Center, Charmo Center for Scientific Research & Consulting, Charmo University, Chamchamal, Sulaimani, KRG, Iraq. E-mail: soran.sedeeq@charmouniversity.org

Received: 12-10-2022 Accepted: 24-12-2022 Published: 15-01-2023

This study presents a transformers-based model using the vocabulary dictionary concept. We collect the parallel corpora in the language and merge them to be a large corpus for training and report the results. The resources used for the task include the Tanzil corpus [5], TED corpus [6], KurdNet–the Kurdish wordnet [7], and the Auta corpus [8].

2. RELATED WORKS

Few studies have addressed the Kurdish language in the Machine Translation (MT) domain. The Apertium project is the first machine translation system for both Sorani and Kurmanji. The Apertium uses rules-based machine translation, which has developed various tools and resources for the Kurdish language, such as bilingual and morphological dictionaries, structural transfer rules, and grammar [4]. InKurdish1 is another attempt to construct a machine translation model for Kurdish. The system applies dictionary-based methods for translation. According to Taher *et al.* (2017), this method is ineffective in translating lengthy and idiomatic sentences. Finally, Ahemdi and Mansoud (2020) attempted to translate Kurdish texts using neural machine translation [4]. Their work was based on collecting the parallel datasets in the Kurdish language. They used different tokenization techniques for training the dataset. They eventually reported the Bilingual evaluation understudy (BLEU) achieved using each tokenizer. Regarding other low-resourced languages worldwide, Abbott and Martinus (2018) employed transformer models to translate texts from English to Setswana using the parallel Autshumato dataset [9]. The outcome of their work indicated that the transformer outperforms previous methods by 5.33 BLEU points. Moreover, Przystupa and Abdul-Mageed (2019) used transformer models with back-translation. Their results demonstrate that transformer models translate texts between Spanish–Portuguese and Czech–Polish [10]. Tapo *et al.* (2019) used neural machine translation to translate texts from Barbara's language to English and French. Their work mainly concentrated on the challenges when performing neural machine translation on a low-resourced language such as Barbara [11].

2.1. Dataset

We used a collection of four parallel datasets. The first one is Tanzil, a group of Quran translations compiled by the Tanzil project8. The corpus has one Sorani translation aligned with 11 translations, totaling 92,354 parallel texts with 3.15M vocabularies on the Sorani Kurdish side and 2.36M on the English side. The corpus is available in translation memory exchange (TMX), where aligned verses are offered [4].

The second corpus, the TED corpus [6], is the collection of subtitles from TED Talks, a sequence of top-notch talks on different genres, "Technology, entertainment, and design." The only Kurdish dialect for which these subtitles are translated is the Sorani dialect. Even though there are only 2358 parallel sentences, the TED collection has translations in a broader, more comprehensive range of subjects than Tanzil.

The third corpus is WordNet [12], a lexical-semantic tool exploited for various Natural Language Processing (NLP) tasks like information extraction and word disambiguation. WordNet offers concise definitions and uses examples for groupings of synonyms, also known as synsets, in addition to semantic links like synonymy, hyponymy, and meronymy. Kurdish WordNet [7] is based on a semi-automatic technique that focuses on creating a Kurdish alignment for base concepts, a critical subset of WordNet's central meanings. Four thousand six hundred and sixty-three definitions directly translated from the Princeton WordNet are included in the most recent version of KurdNet (version 3.0). We included this resource despite having fewer translated purposes than necessary for machine translation because it covers more domains.

The final corpus is Auta, comprising 229,222 pairs of physically aligned translations [8]. The corpus is gathered from different text genres and domains to construct more solid and real-world machine translation applications. The researchers built this corpus and published a portion of this corpus available to promote study in this area, which contains 100.000 normalized and cleaned texts ready to be experimented with using the trendy machine learning models (Table 1).

2.2. Transformer's Model Architecture

Most neural machine translation models follow an encoder-decoder structure [13]. The encoder consists of six identical layers with a multi-head self-attention mechanism and position-wise sublayers. These layers are fully connected to feed-forward networks. The encoder aims to map an input sequence of symbol representations starting from (x1;::; xn) to a sequence of continuous representations, which is z = (z1;::; zn) [14].

Table 1: Size of each Kurdish–English corpus

No	Corpus	Language	Size
1	Tanzil	Kurdish–English	92.354 texts
2	Ted	Kurdish–English	2.358 texts
3	KurdNet	Kurdish–English	4.663 texts
4	Auta	Kurdish–English	100.000 texts
Total			199.375 texts

Similarly, the decoder is constituted of a stack of six identical layers. Encoder layers consist of two sublayers each, and the decoder adds the third sublayer to carry out multi-head attention around the encoder`s output. In the same way as the encoder, layer normalization uses residual connections around each sub-layer. The self-attention sub-layer in the decoder stack has been adjusted to block positions from attending to the following positions, as shown in Fig. 1. The goal of the decoder is to produce a sign output sequence (y1;:; ym) one element at a time [14].

Moreover, the model uses the attention function to map a query and a set of key-value pairs to an output, where the question, keys, values, and production are all vectors. The sum of the weight values calculates the result. A compatibility function between the query and the relevant key determines each value’s weight.

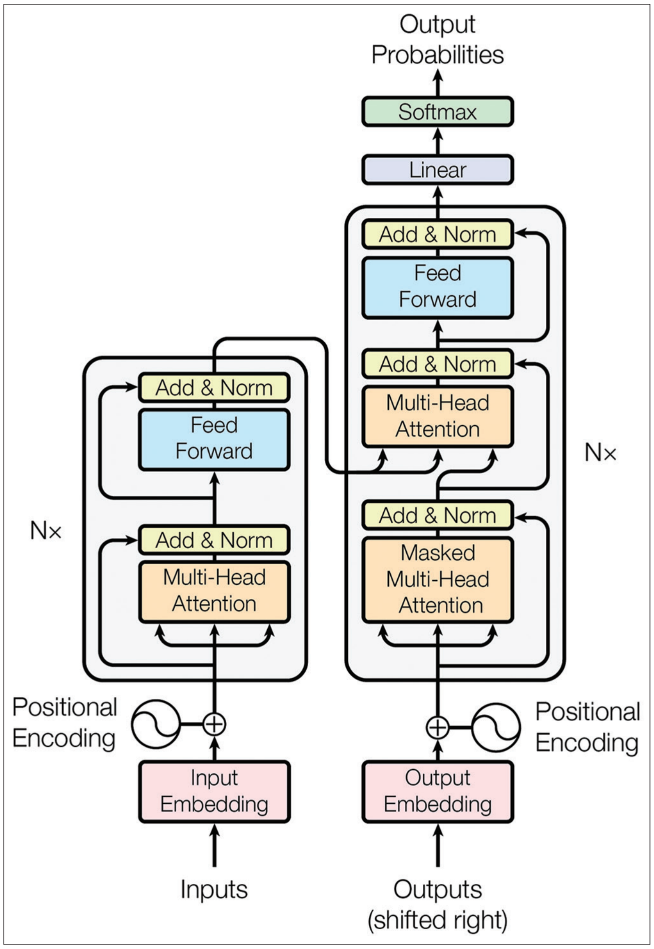


Fig. 1. The transformer model architecture [15].

As is shown, the transformer operates multi-head attention on three different stages:

1. Encoder-decoder attention lets every decoder position focus on every input post.
2. The encoder has layers for the self-attention. All the keys, values, and queries in a self-attention layer originate from the same source, in this case, the encoder’s output from the previous layer.
3. The decoder’s self-attention layers enable each location to pay attention to all postings below and above.

A fully connected feed-forward network is implemented to each position separately and uniformly in each layer of the encoder and decoder. Two linear transformations and a ReLU (Rectified Linear Unit) activation make up this process [15].

The decoder output is transformed to project next-token probabilities using the SoftMax function. The embedding is utilized to convert the input and output tokens to vectors of the dimension model.

The positional encodings to the input embedding at the bottoms of the encoder and decoder stacks. Since the positional encodings and the embeddings share the same dimension model, both can be added. Positional encodings come in a variety of discovered and fixed forms [16].

2.3. The Proposed Model Architecture

The decoder and encoder for the proposed transformer-based Neural Machine Translation (NMT) have a stack of six layers, as shown in Fig. 2. Every layer has two sublayers: The position-wise feed-forward sub-layer and the multi-head attention sub-layer (FFN). The encoder and decoder in the proposed Transformer NMT model architecture for Kurdish texts produce variable-length sequences using an attention model and feed-forward net. Multi-head attention is the foundation for how attention operates across multiple tiers. The mapping of an input sequence of symbol representations, X = (x1, x2..., xnenc)T to an intermediate vector. Given the intermediate vector, the decoder generates the output sequence (target sentence) Y = (y1, y2..., yndec)T. The convolutional or recurrent structures are absent from the transformer design. At the first layer of both the encoder and the decoder, the positional encodings computed by the Equations below are summed to the input embeddings.

1) PE(pos, 2i) = sin(pos100002i/dmodel)
2) PE(pos, 2i+1) = cos(pos100002i/dmodel)

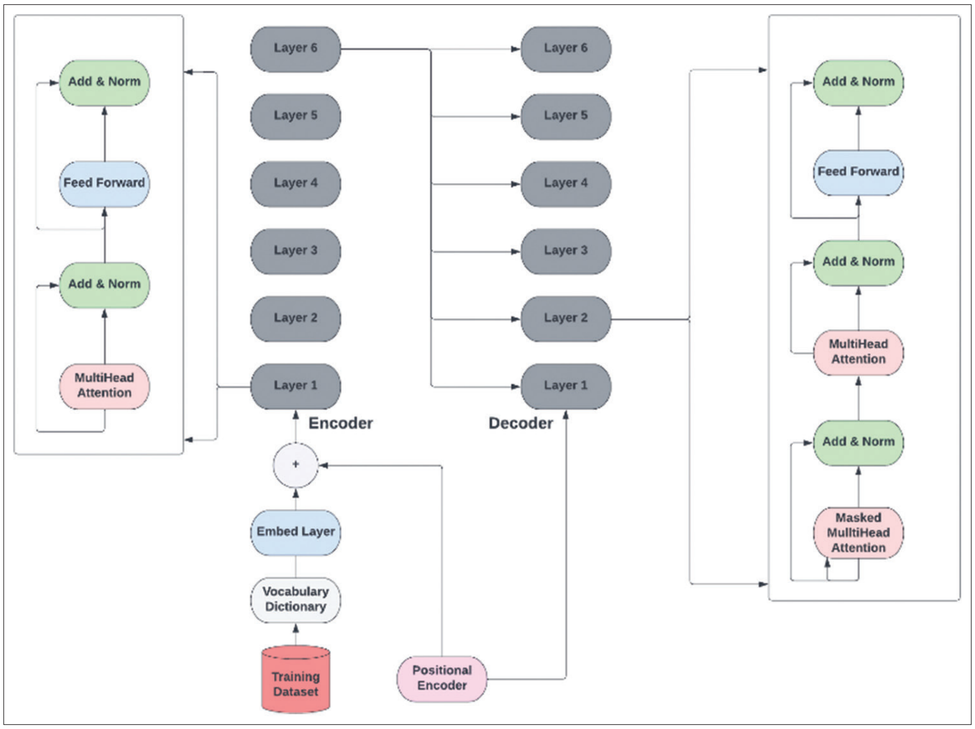


Fig. 2. The Proposed model architecture.

Where pos stands for position, i is considered the dimension, and d is the dimension of the intermediate representation. Every encoding layer has a position-wise feed-forward sub-layer and a multi-head attention sub-layer. A residual connection method [17] and a layer normalization unit (LayerNorm) [13] are used around each sub-layer to facilitate training and enhance performance. In contrast to the encoder, every layer of the decoder has three sub-layers: a multi-head attention sublayer, a position-wise feed-forward sub-layer, and so on. Encoder-decoder multi-head attention sub-layer is inserted in between them.

3. METHODOLOGY

The proposed model uses the concept word dictionary inside the dataset to find the equivalence meaning of each word. Therefore, in the preprocessing stage, we only tokenized the cleaned texts, which converted the sentences into lists of words. Following that, we converted them into an extensive dictionary of words which has Kurdish words and their English meanings. Next, we fed the dictionary to our proposed transformer's model. As shown in Fig. 3. We used the batch size of 20 and trained on 100 epochs.

At first, we tried to train the model on the central processing unit (CPU); since the amount of data was huge for the CPU, the model trained for days without providing any results, and numerous ram crashes forced the computer to reboot and restart the process again. However, we tried to train the model for one epoch and compare its result with graphics processing unit (GPU). As it is shown in Table 2.

As shown in Table 2, training one epoch on the CPU lasted 3 h and 37 min, while it lasted <5 min for GPU. Because 100 epochs are enormous to be trained on CPU and to avoid Ram crashes, we trained the model on Google Colab Pro, a monthly subscription program that gives you higher Ram and GPU. The whole training and test process lasted 5 h. The complete code and the training program are publicly available at/https://github.com/mbrow309/MachineTranslationUsingTransformers/blob/master/KurdishMTTransformers.ipynb/.

4. RESULTS AND DISCUSSION

We train the system on 100 epochs since introducing the MT module at higher values will help guarantee a good BLEU score. Neural networks are usually trained over several

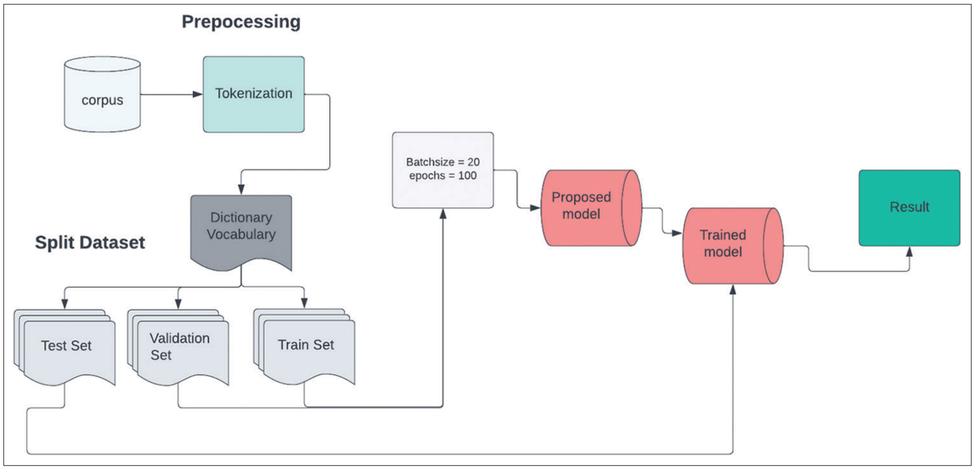


Fig. 3. An outline of the methodology.

Table 2: The difference between training on GPU and CPU

Machine type	Loss	BLEU	Time
CPU	5.140	0.0183	3 h and 37 min
GPU	5.109	0.0191	4 min

GPU: Graphics processing unit, CPU: Central processing unit, BLEU: Bilingual evaluation understudy

Table 3: Samples of the produced translation texts

Kurdish	من تۆم خۆشدەویت
English	I love you, and I love you.
Kurdish	من دەچم بۆ بازاڕ
English	I am going to go to the marketing game.
Kurdish	ڕۆژێک لە ڕۆژان نەخۆشیەکی ترسناک بوو بە هەڕەشە
English	Once upon a time, there was a dread disease.
Kurdish	ئێوە لە کوێ بوون
English	Where were you, like yesterday?

epochs. Epochs refer to cycles through a training dataset [18]. It is important to note that we tried to train the module on each dataset, and due to the low amount of datasets, the module yielded a significantly lower BLEU score. However, merging the datasets did a perfect job. As shown in Fig. 4, the amount of BLEU improves significantly per 10 epochs and finally reaches the ideal score.

We fed the module some unseen texts to translate. Overall, the module did an excellent job of translating the texts. Below are samples of translated texts shown by the module.

The model does a relatively good job of translating unseen texts. Even though the translation results from Table 3 show some cases of word repetition and some cases of producing ungrammatical sentences, particularly in the final test example. The issue is substantially related to having a low amount of data. Therefore, if the model is trained on much larger datasets, the translation results would be more accurate and flawless.

In the next phase of our work, we intend to investigate the ergative case of our model by feeding it examples that have ergative and compare our model`s translation with the latest Google translation for the Kurdish language. In Kurdish, the word order is subject-object-verb with tense-aspect-modality markings [19]. As a split-ergative language, Sorani Kurdish marks transitive verbs in the past tenses differently from nominative verbs [20]. For ergative-absolute alignment, Sorani Kurdish uses different pronominal enclitics [4]. To clarify further, we have included a few examples in Sorani Kurdish below. The bold suffix is used for patient marking in Example 1 in the past tense, which uses the pronominal enclitic = man as an agentive marker.

1. Kurdish/منداڵەکانمان هێنان
Transcribe/mndalakanman hênan.

Translation/we brought the children.

2. Kurdish/هێنامانن
Transcribe/hêna**man**in

Translation/we brought them.

3. Kurdish/دەچنە باخەکامان

Transcribe/deçine baxakaman

Translation/they are going to our garden.

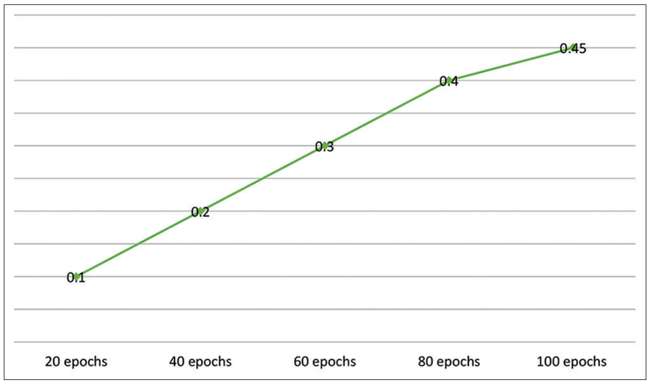


Fig. 4. The value bilingual evaluation understudy value per 20 epochs.

Table 4: Comparison between translating ergative sentences using Google translator and transformer-based model

Google translator	Our model	Kurdish text
Sutandmian	They burnt me	سوتاندمیان
brought it	A two into difficulties	هێنایانی
They took them to the market.	They took me to the market.	بردمیان بۆ بازاڕ
They brought them home.	They took me home	هێنامیان بۆ ماڵ
I saw them in the market.	I saw them at the market.	بینیانم لە بازاڕ
I love you.	I love you.	من تۆم خۆشدەوێت

As shown, the pronominal (man) is translated differently in different examples. In examples 1 and 2, "man" was the subject, and its equivalence is "we." While in example 3, it functions as a possessive pronoun, and it is "our." During machine translation, this creates significant issues when the model tries to align the two languages.

The examples in Table 4 show that our model performs well in the ergative situation for Kurdish texts. According to the results, the Google translator faces issues when the sentence contains the pronominal enclitics (m), and it functions as the object of the sentence. This is because our corpus includes many natural language texts that include such pronominal pronouns, particularly in the Tanzil corpus. Thus, our model would easily detect the pronominal enclitics and their alignment inside the texts.

5. CONCLUSION

The transformer model is a unique and highly functional model to translate texts from one language to another. Undoubtedly, the Kurdish language suffers from a lack of resources, particularly in the field of NLP. The lack of a translation model is also part of the problem. The work undertaken in this paper demonstrates that the Kurdish language responds well to the newly developed and proposed neural machine translation model. It is worth noting that the existence of large corpora with more than 1 million data can actively work well and improve the model's score to near-perfect translation. Fortunately, the results acquired from this work can open many gates for the future researchers to dive deeply into the transformer model and modified in a way that can work specifically for the language. Finally, the transformer model's layers remain intact, and the training and process started this way as the model modification, particularly on the layers left for future researchers.

REFERENCES

[1] S. Tripathi and J. K. Sarkhel. "Approaches to machine translation". *Annals of Library and Information Studies*, vol. 57, pp. 383-393, 2010.

[2] P. Koehn. "*Statistical Machine Translation*". Cambridge University Press, Cambridge. 2009.

[3] L. Bentivogli, A. Bisazza, M. Cettolo and M. Federicoa. "Neural versus phrase-based mt quality: An in-depth analysis on english-german and english-french". *Computer Speech and Language,* vol. 49, pp. 52-70, 2019.

[4] S. Ahmadi and M. Masoud. "Towards Machine Translation for the Kurdish Language". *arXiv preprint arXiv:2010.06041,* 2020.

[5] J. Tiedemann. "Parallel data, tools and interfaces in OPUS". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12*). European Language Resources Association (ELRA), Istanbul, Turkey. pp. 2214-2218, 2012.

[6] M. Cettolo, C. Girardi and M. Federico. "Wit3: Web inventory of transcribed and translated talks". In: *Conference of European Association for Machine Translation*. 2012.

[7] P. Aliabadi, M. S. Ahmadi, S. Salavati and K. S. Esmaili. "Towards building kurdnet, the kurdish wordnet". In: *Proceedings of the Seventh Global Wordnet Conference*. University of Tartu Press, Tartu, Estonia. 2014.

[8] Z. Amini, M. Mohammadamini, H. Hosseini, M. Mansouri and D. Jaff. "Central Kurdish Machine Translation: First Large Scale Parallel Corpus and Experiments". *arXiv preprint arXiv:2106.09325,* 2021.

[9] L. Martinus and J. Z. Abbott. "A Focus on Neural Machine Translation for African Languages". *arXiv preprint arXiv:1906.05685,* 2019.

[10] M. Przystupa and M. Abdul-Mageed. "Neural machine translation of low-resource and similar languages with backtranslation". In: *Proceedings of the Fourth Conference on Machine Translation*. vol. 3. Association for Computational Linguistics, Florence, Italy. 2019.

[11] A. A. Tapo, B. Coulibaly, S. Diarra, C. Homan, J. Kreutzer, S. Luger, A. Nagashima, M. Zampieri and M. Leventhal. "Neural Machine Translation for Extremely Low-Resource African Languages: A Case Study on Bambara". *arXiv preprint arXiv:2011.05284,* 2019.

[12] G. A. Miller. "*WordNet: An Electronic Lexical Database*". MIT Press,

Massachusetts, United States. 1998.

[13] J. L. Ba, J. R. Kiros and G. E. Hinton. "Layer Normalization". *arXiv preprint arXiv:1607.06450,* 2016.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin. "Attention is all you need". In: *Conference on Advances in Neural Information Processing Systems.* 2017.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. "Dropout: A simple way to prevent neural networks from overfitting". *The Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[16] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y. N. Dauphin. "Convolutional sequence to sequence learning". In: *Proceedings of the 34th International Conference on Machine Learning* (PMLR). 2017.

[17] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[18] L. N. Smith. "Cyclical learning rates for training neural networks". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Santa Rosa, CA, USA. 2017.

[19] G. Hai and Y. Matras. "Kurdish linguistics: A brief overview". *STUF-Language Typology and Universals*, vol. 55, pp. 3-14, 2002.

[20] M. R. Manzini, L. M. Savoia and L. Franco. "Ergative case, aspect and person splits: Two case studies". *Acta Linguistica Hungarica*, vol. 52, pp. 297-351, 2015.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

ECG Signal Recognition Based on Lookup Table and Neural Networks

Muzhir Shaban Al-Ani

University of Human Development, College of Science and Technology, Department of Information Technology, Sulaymaniyah, KRG, Iraq

ABSTRACT

Electrocardiograph (ECG) signals are very important part in diagnosis healthcare the heart diseases. The implemented ECG signals recognition system consists hardware devices, software algorithm and network connection. An ECG is a non-invasive way to help diagnose many common heart problems. A health-care provider can use an ECG to recognize irregular heartbeats, blocked or narrowed arteries in the heart, whether you have ever had a heart attack, and the quality of certain heart disease treatments. The main part of the software algorithm including the recognition of ECG signals parameters such as P-QRST. Since the voltages at which handheld ECG equipment operate are shrinking, signal processing has become an important challenge. The implemented ECG signal recognition approach based on both lookup table and neural networks techniques. In this approach, the extracted ECG features are compared with the stored features to recognize the heart diseases of the received ECG features. The introduction of neural network technology added new benefits to the system implementing the learning and training process.

Index Terms: Electrocardiograph signals, P-QRS, Healthcare, Heart diseases.

1. INTRODUCTION

Patients suffering from heart diseases need continuous healthcare especially for ECG monitoring and recognition to avoid dangerous of heart failure [1]. The reduction of heart attack depends on the fast identification of abnormal cardiac rhythms [2]. ECG is an effective diagnostic technique which is widely used by cardiologists [3]. ECG are electrical signals of the heart recorded by electrodes fixed on patient body [4]. ECG signals provide useful information about the rhythm and the operation of the heart. Heart beats extracted from ECG signals can be categorized into classes that are: Normal, atrial premature, and ventricular escape beats [5].

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp22-31

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Electrocardiographs are recorded by electrocardiograms that are very important for healthcare diseases [6]. These devices record electrical signal picked up by electrodes attached to certain parts of the patient body [7]. The signals recorded by the electrocardiograms at any moment are the sum of the all signals passing in cells throughout the heart [8]. Electrocardiogram consists of 12 leads which indicated 12 electrical views of the heart [9]. The first six leads represent the frontal plane leads; I, II, III, V_R, V_L and V_F. Leads I, II, and III are the standard leads and are find by [10]:

$$I = V_L - V_R \quad (1)$$

$$II = V_F - V_R \quad (2)$$

$$III = V_F - V_L \quad (3)$$

The other six leads are in the front of the heart; V_1, V_2, V_3, V_4, V_5, and V_6, these are recorded by the six electrodes placed on the chest of the patient [11].

Corresponding author's e-mail: muzhir.al-ani@uhd.edu.iq

Received: 22-10-2022 Accepted: 08-01-2023 Published: 21-01-2023

Electrocardiograms are concentrated on all issues associated with diseases of heart attack patients that used directly with clinical [12]. Recently, a reliable and automatic analysis and segmentation of ECG signals are required for health-care environments [13]. Computer based methods are suitable for processing and analyzing of ECG signals [14]. Artificial neural network techniques are used for analyzing different types of signals and tasks related to heart diseases [15]. Most of these tasks are associated to the detection of irregular heartbeats and irregular in recording process [16]. A back propagation neural network may apply in training stage to give a powerful pattern recognition algorithm [17].

Electrocardiograph (ECG) signals are very important part in diagnosis healthcare the heart diseases. The implemented ECG signals recognition system consists hardware devices, software algorithm and network connection. An ECG is a non-invasive way to help diagnose many common heart problems. A health-care provider can use an ECG to recognize irregular heartbeats, blocked or narrowed arteries in the heart, whether you have ever had a heart attack, and the quality of certain heart disease treatments. The main part of the software algorithm including the recognition of ECG signals parameters such as P-QRST. Since the voltages at which handheld ECG equipment operate are shrinking, signal processing has become an important challenge. The implemented ECG signal recognition approach based on both lookup table and neural networks techniques. In this approach, the extracted ECG features are compared with the stored features to recognize the heart diseases of the received ECG features.

2. ECG SIGNALS

Heart diseases are the well-known disease that affects humans worldwide [18]. Yearly millions of people die or suffered from heart attacks [19]. Early detection and treatment of heart diseases can prevent such events [20]. This would improve the quality of life and slow the events of heart failure [21]. The main benefit of the diagnosis is to record the ECG of the patient [22]. An ECG record is a non-invasive diagnostic tool used for the assessment of a patient heart condition [23]. The extraction of ECG features and combined that with the heart rate, these can lead to a fairly accurate and fast diagnosis [24].

Bioelectrical signals represent human different organs electrical activities and ECG signals are the important signals among bioelectrical signals that represent heart electrical activity [25]. Deviation or distortion in any part of ECG that is called arrhythmia can illustrate a specific heart disease [26]. The investigation of the ECG has been extensively used for diagnosing many cardiac diseases [27]. The ECG is a realistic record of the direction and magnitude of the electrical commotion that is generated by depolarization and repolarization of the atria and ventricles [28].

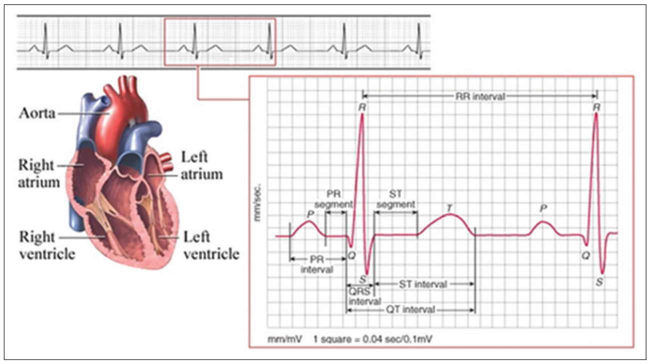


Fig. 1. Representation of electrocardiograph signals.

One cardiac cycle in an ECG signal consists of the P-QRS-T waves as shown in Fig. 1 [29,30]. The majority of the clinically useful information in the ECG is originated in the intervals and amplitudes defined by its features (characteristic wave peaks and time durations) [31,32]. ECG is essentially responsible for patient monitoring and diagnosis [33].

Normal rhythm produces four entities; P wave, QRS complex, T wave, and U wave in which each have a fairly unique pattern as shown in Fig. 1 [34,35]:

- P-Wave - Represents the movement of an electric wave from the sino atrial (SA) node and causes depolarization of the left and right atria.
- P-R segment - Represents the pause in electrical activity caused by a delay in conduction of electrical current in the atrioventricular (AV) node to allow blood to flow from the atria to the ventricles before ventricular contraction happen.
- QRS complex - Represents the electrical activity from the beginning of the Q wave to the end of the S wave and the complete depolarization of the ventricles, resulting to ventricular contraction and ejection of blood into the aorta and pulmonary arteries.
- S-T segment - Represents the pause in electrical activity after complete depolarization of the ventricles to allow blood to flow out of the ventricles before ventricular relaxation begins and the heart to fill the next contraction.
- Wave T - Represents the repolarization of the ventricles.

- Wave U - Represents the repolarization of the papillary muscle.

3. LITERATURE REVIEWS

Mohammed *et al.,* presented an ECG compression algorithm based on the optimal selection of wavelet filters and threshold levels in different sub bands that allow a maximum reduction of the volume of data guaranteeing the quality of the reconstruction. The proposed algorithm begins by segmenting the ECG signal into frames; where each image is decomposed into sub bands m by optimized wavelet filters. The resulting wavelet coefficients are limited and those having absolute values below the thresholds specified in all sub bands are eliminated and the remaining coefficients are properly encoded with a modified version of the run coding scheme [36].

Reza *et al.,* proposed compressed detection procedure and the collaboration detection matrix approach that used to provide a robust ultra-light energy focus for normal and abnormal ECG signals. The simulation results based on two proposed algorithms illustrate a 15% increase in signal to noise ratio and a good quality level for the degree of inconsistency between random and scatter matrices. The results of the simulation also confirmed that the Toeplitz binary matrix offered the best SNR performance and compression with the highest energy efficiency for the random array detection [37].

Ann and Andrés implemented an approach to classify multivariate ECG signals as a function of analyzing discriminant and wavelets. They used variants of multiscale wavelets and wave correlations to distinguish multivariate ECG signal models based on the variability of the individual components of each ECG signal and the relationships between each pair of these components. Using the results from other ECG classification studies in the literature as references that demonstrated this approach to 12-lead ECG signals from a particular database compares favorably [38].

Vafaie, *et al.,* presented a new classification method to classify ECG signals more precisely based on the dynamic model of the ECG signal. The proposed method is constructed a diffuse classifier and its simulation results indicate that this classifier can separate the ECG with an accuracy of 93.34%. To further improve the performance of this classifier, the genetic algorithm is applied when the accuracy of the prediction increases to 98.67%. This method increased the precision of the ECG classification for a more accurate detection of the arrhythmia [39].

Kamal and Nader, realized a practical means to synthesize and filter of ECG signal in the presence of four types of interference signals: first, from electrical networks with a fundamental frequency of 50 Hz, second, those resulting from breathing, with a frequency range 0.05–0.5 Hz, third musical signals with a frequency of 25 Hz and fourth white noise presented in the ECG signal band. This was accomplished by implementing a multiband digital filter (seven bands) of the finite impulse response multiband least square type using a programmable digital apparatus, which was placed on an education and development board [40].

Farideh *et al.,* explored combined discriminative ability of ECG/R signals in automatic staging. Basically, this approach classified that the wakefulness of slow wave sleep and REM sleep was classified using a vector support machine fed with a set of functions extracted from characteristics of 34 features and characteristics of 45 features. First part has produced a reasonable discriminatory capacity, while the second part has considerably improved the rating and the best results were obtained using third approach. We then improved the support vector machine classifier with the recursive feature elimination method. The results of the classification were improved with 35 of the 45 features [41].

Shirin and Behbood classified a patient's ECG cardiac beats into five types of cardiac beats as recommended by AAMI using an artificial neural network. This approach used block based on the neural network as a classifier. This approach created from a set of two dimensional blocks that are connected to each other. The internal structure of each block depends on the number of incoming and outgoing signals. The overall construction of the network was determined by the movement of signals through the network blocks. The network structure and weights are optimized using the particle swarm optimization approach [42].

Prakash and Shashwati, proposed an approach that attempts to reduce unwanted signals using a Minorization-Maximization method to optimize total signal variation. The unsuccessful signal is then segmented using the bottom-up approach. The obtained results show a significant improvement in the signal-to-noise ratio and the successful segmentation of the ECG signal sections. The extension of the heel depends on the smoothing parameter of Lamda. As this approach was implemented for complete signal, then only 18 dB of signal to noise ratio was achieved [43].

Aleksandar and Marjan, focused on a new algorithm for the digital filtering of an electrocardiogram signal received by stationary and non-stationary sensors. The basic idea of digital processing of the electrocardiogram signal is to extract the heartbeat frequencies that are normal in the range between 50 and 200 beats/min. The frequency of the extracted heart rate is irregular if the rate increases or decreases and serves as evidence for the diagnosis of a complex physiological state. The environment can generate a lot of noise, including the supply of electrical energy, breathing, physical movements, and muscles [44].

Kumar *et al.,* proposed an automated diagnosis of coronary artery disease using electrocardiogram signals. Flexible Analytical Wavelet Transform technology is used to break down electrocardiogram effects. The Cross Information Potential parameter is calculated from the actual values of the Flexible Analytical Wavelet Transform decomposition detail coefficients. For diagnosis of coronary artery disease subjects, the mean value of the Cross Information Potential parameter is higher in the comparison toner subjects. The statistical test is applied to check the discrimination capacity of the extracted functionalities. In addition, the functionality is fed to the least squares support vector machine for sorting. The classification accuracy is calculated at each decomposition level from the first decomposition level [45].

Al-Ani, explained that ECG waveform is an important process for determining the function of the heart, so it is useful to know the types of heart disease. The ECG chart gives a lot of information that is converted into an electrical signal containing the basic values in terms of amplitude and duration. The main problem that arises in this measurement is the confusion between normal and abnormal layout, in addition to certain cases where the P-QRS-T waveform overlaps. The purpose of this research is to provide an effective approach to measure all parts of the P-QRS-T waveform to give the right decision for heart function. The proposed approach depends on the classifier operation that based mainly on the features extracted from electrocardiograph waveform that achieved from exact baseline detection [46].

Nallikuzhy and Dandapat, explored an efficient technique to improve a low resolution ECG by merging fragmented coding and the learning model of the common dictionary. An enhance model is applied on low resolution ECG using previously learned model in order to obtain a high resolution full estimate of 12-lead ECG. This approach was applied based on the dictionary in which the common dictionary contains high and low resolution dictionaries regarding to the high and low resolution ECG and is learned simultaneously. Similar fragmented representation for high and low resolution ECGs was generated using Joint dictionary learning. Mapping between the scattered coefficients of the high and low resolution ECGs was also learned [47].

Han and Shi presented an efficient method of detection and localization of myocardial infarction that combines a multi-lead residual neural network structure (ML-ResNet) with three residual blocks and a function fused by 12-lead ECG recordings. A single network of characteristic branches was formed to automatically learn representative characteristics of different levels between different layers, which exploit the local characteristics of the ECG to characterize the representation of spatial information. Then, all the main features are merged as global features. To evaluate the generalization of the proposed method and clinical utility, two schemes are used that include the intra-patient scheme and the inter-patient scheme. The obtained results indicated a high performance of accuracy and sensitivity [48].

Abdulla and Al-Ani, implemented a review study classification for ECG Signal. This work aimed to investigate and review the use of classification methods that have been used recently, such as the artificial neural network, the convolutional neural network, discrete wavelets transform, support vector machine and K-Nearest Neighbor. Effective comparisons are presented in the result in terms of classification methods, feature extraction technique, data set, contribution, and some other aspects. The result also shows that convolutional neural network has been used more widely for ECG classification as it can achieve higher accuracy compared to other approaches [49].

Abdulla and Al-Ani, explained an automatic ECG classification system which is difficult to detect, especially in manual analysis. An accurate classification and monitoring ECG system was proposed using the implementation of convolutional neural networks and long-short term memory. Learned features are captured from the CNN model and passed to the LSTM model. The output of the CNN-LSTM model demonstrated superior performance compared to several of the more advanced ones cited in the results section. The proposed models are evaluated on the MIT-BIH arrhythmia and PTB diagnostics datasets. A high accuracy rate of 98.66% in the classification of myocardial infarction was obtained [50].

4. METHODOLOGY

The methodology of this approach is divided into three parts: ECG signals recognition approach, ECG feature extraction

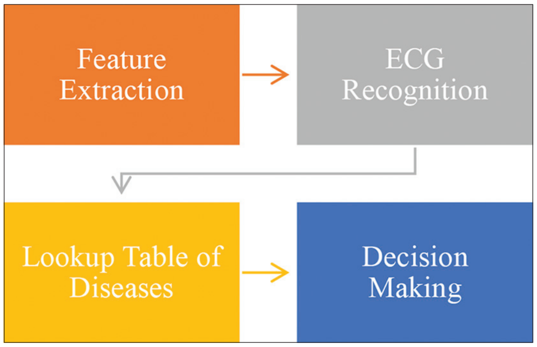


Fig. 2. Electrocardiograph recognition approach.

and neural network architecture. In addition, the used data are images selected with different heart diseases.

The main objective of introducing forward propagation neural networks in this work is to determine the main component values of the ECG signal, which are seven values (QRS complex, QT interval, QTCB wave, PR interval, P wave, RR interval and PP interval) and to compare that with the table that carries the standard values. Depend on this comparison, it is possible to make an accurate decision about whether the ECG signal is normal or abnormal.

4.1. ECG Signals Recognition Approach

The ECG signals recognition approach is implemented through the following stages (Fig. 2):

- Feature extraction stage in which ECG signals parameters (amplitude and time interval) will be extracted from the electrodes.
- ECG recognition stage in which the extracted parameters of ECG are applied through neural network that specified the diseases associated with these parameters.
- Lookup table stage in which the constructed lookup table is associated with the list of specified heart diseases.
- Decision making stage in which take the decision of which type of heart diseases are related.

4.2. ECG Feature Extraction

Tracing of ECG signal on the special recognition is very important to extract the values of the direct parameters. The main advantage of ECG feature extraction operation is to generate a small set of features that achieve the ECG signal. ECG feature extraction operation is implemented through many steps as shown in Fig. 3. The first step is preprocessing in which ECG graph will be cleaned and resized. The second step focusing on thinning filter in which the ECG signal will be better quality, in addition this step will eliminate the scattering pixels around the original signal. The third step concentrates on edge detection that detects the original ECG signal. In addition, this indicates the duration and amplitude of each ECG signal part. Fourth step is to calculate the required parameters of ECG signal that related on duration and amplitude of each part of ECG signal.

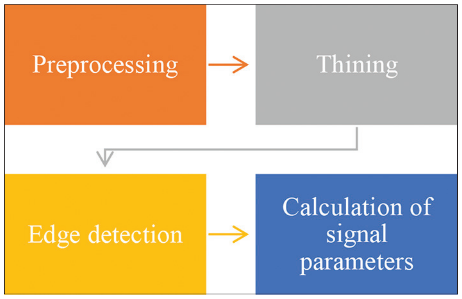


Fig. 3. Electrocardiograph feature extraction.

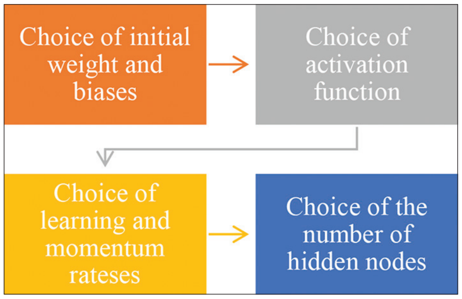

Choice of initial weight and biases
Choice of activation function
Choice of learning and momentum rateses
Choice of the number of hidden nodes

Fig. 4. Design of neural network architecture.

4.3. Neural Network Architecture

Proper neural network architecture is used to be efficient and work in a wide range of conditions. It is necessary to choose network parameters such that the obtained ECG system is acceptable for theoretical and practical settings. Furthermore, this neural network is an application-oriented system and the design is done with the selection of the network architecture. In this case, many parameters are selected as below (Fig. 4):

- Choice of initial weight and biases: The choice of initial weight will influence how quickly the system coverage? The values of the initial weights must not be too large or too small to avoid out of region condition. The weights and biases of ECG network in learning phase are initialized randomly between -0.5 and 0.5.
- Choice of activation function: The ECG used neural network of sigmoid function which has simple derivative and nonlinear property. The sigmoid range of output lies between zero and one.
- Choice of learning and momentum rate: For low learning rate, the neural network will adjust their weights gradually, but the convergence may be slow, while for high learning

rate the neural network has big changes that are not desirable in a trained network. The network consists of 5 nodes in input layer, 80 nodes in hidden layer, and 4 nodes in output layer.

- Choice of the number of hidden nodes: The number of hidden nodes in the hidden layers is varied from 5 nodes to 125 nodes, while keeping the learning rate and momentum rate constant at nominal values (learning rate = 0.7 and momentum rate = 0.9). Backpropagation neural network algorithm is used for ECG system to achieve a balance between correct response to the trained patterns and good responses to new input patterns.

The forward propagation algorithm starts with the presentation of input pattern to the input layer of the network and continues as activation level calculations propagate forward through the hidden layers. Every processing unit (in each successive layer) sums its inputs and applies the sigmoid function to compute its output. Then the output layer of the units produces the output of the network.

Suppose the total input S_j to unit $_j$ is a linear function of the states of units a_i which is equal to the activation levels of the neurons in the previous layer that is connected to unit $_j$ through the weights W_{ji} and the threshold, θ_j of unit $_j$ where:

$$S_j = \sum_i a_i W_{ji} + \theta_j \tag{4}$$

The state of (y) of a unit is a sigmoid function of its total input S.

$$y_i = f\left(S_j\right) = \frac{1}{1 + e^{-s}} \tag{5}$$

The resulting value becomes the activation level of neuron $_j$. once the set of the outputs for a layer is found, it serves as an input to the next layer. This process is repeated layer by layer until the final set of network output is produced.

The backward propagation algorithm indicated by error values and these are calculated for all processing units and the weight changes are calculated for all interconnections. The calculations begin at the output layer and progress backward through the network to the input layer.

The error value is simple to be computed for the output layer and somewhat more complicated for the hidden layers. If unit $_j$ represents the output layer, then its error value is given by:

$$\delta_j = \left(t_j - a_j\right) F'\left(S_j\right) \tag{6}$$

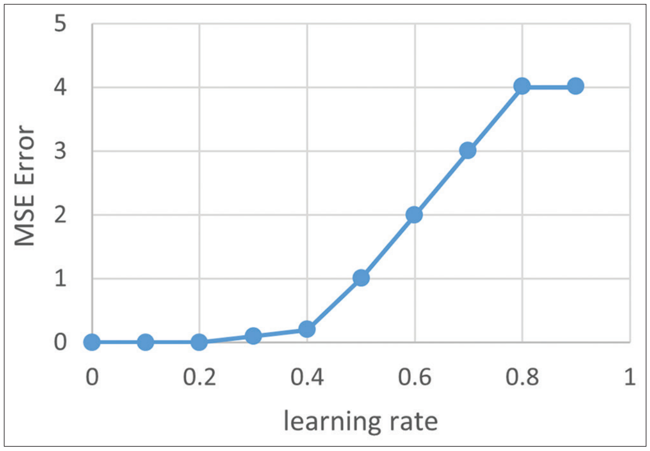


Fig. 5. The relation between learning rate and mean square error.

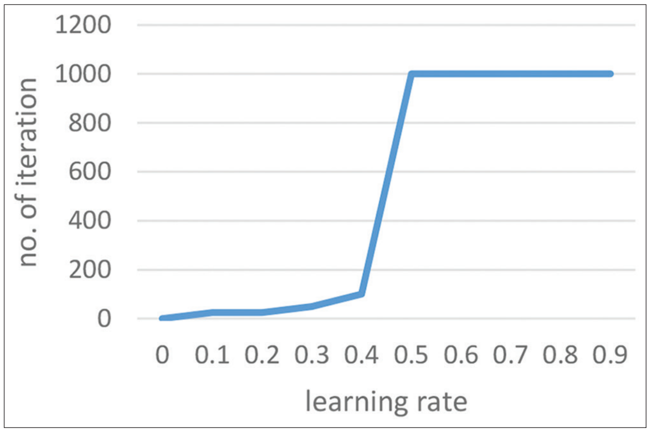


Fig. 6. The relation between learning rate and number of iteration.

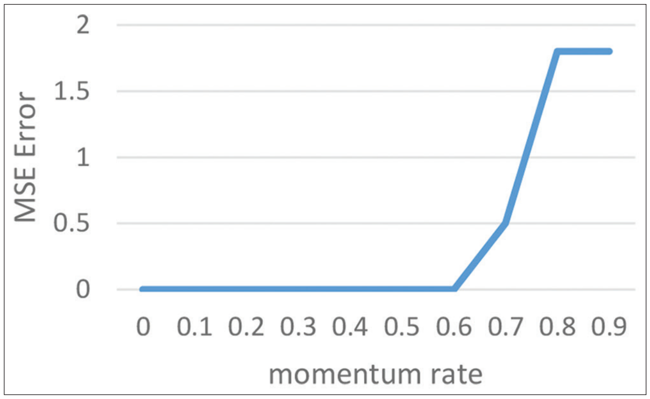


Fig. 7. The relation between momentum and mean square error.

Where:

- t_j is the target value for unit $_j$.
- F' (S_j) is the derivative of the sigmoid function F.
- a_j is the output value for unit j.
- S_j is the weighted sum of inputs to j.

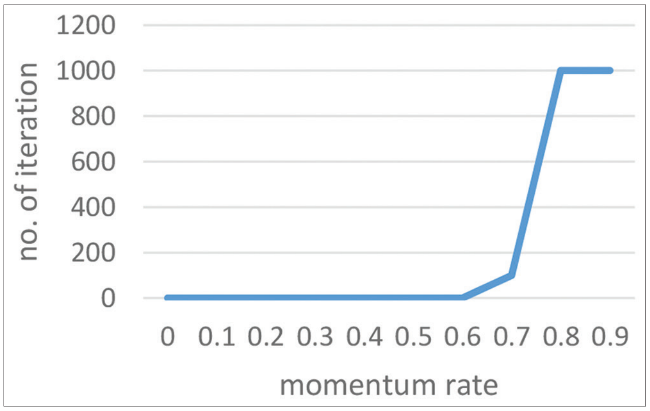


Fig. 8. The relation between momentum and number of iteration.

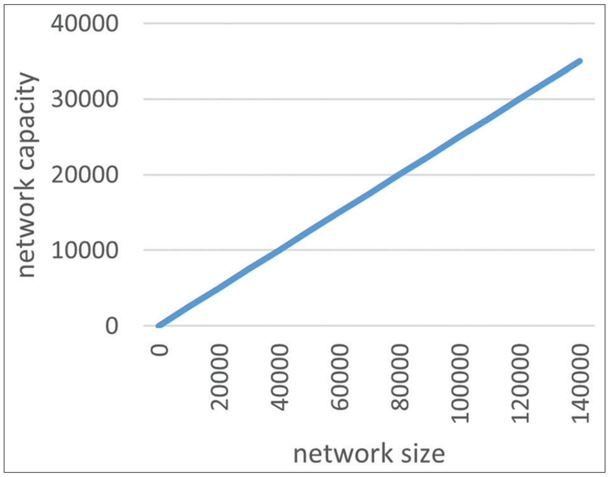


Fig. 9. The relation between network size and network capacity.

5. RESULTS AND DISCUSSION

Fig. 5 gives the relation between learning rate and mean square error (MSE). The learning rate value is laying between zero and one and the commonly used range is in between 0.25 and 0.75. At this active rang of learning rate, the calculated MSE is so small and laying in the range 0.1 and 3.5. Fig. 6 shows the relation between learning rate and the number of iteration. At this figure it is clear that when the learning rate is equal to 0.5, then the number of iteration is 1000 and still saturated at this number of iteration as learning rate increases.

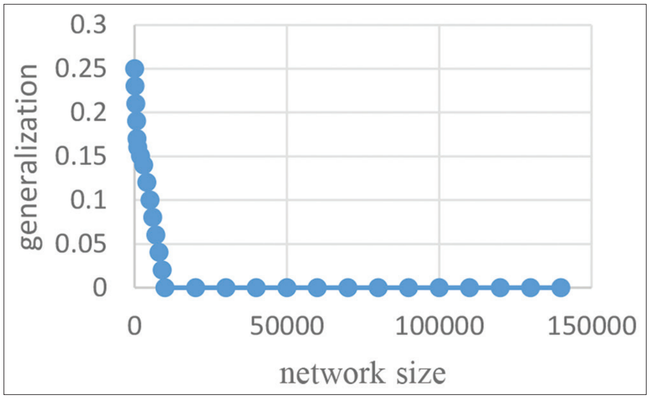


Fig. 10. The relation between network size and generalization.

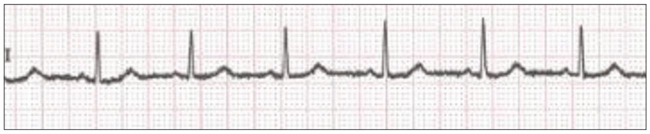


Fig. 11. Normal electrocardiograph signal.

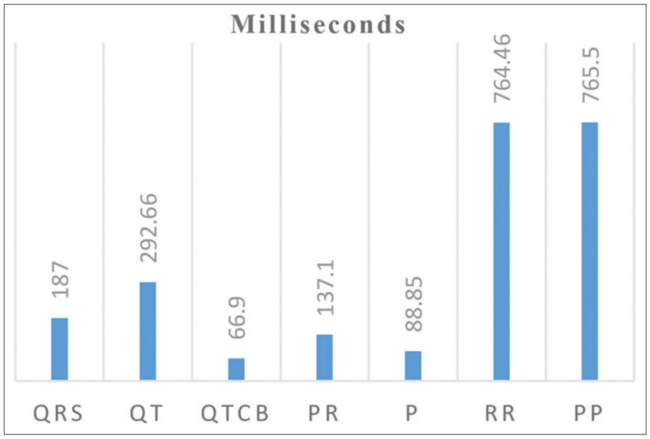


Fig. 12. Final diagnosis of electrocardiograph signals (normal case).

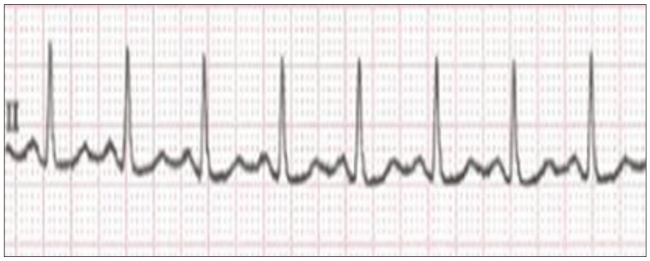


Fig. 13. Sinus tachycardia electrocardiograph signal.

Fig. 7 shows the relation between momentum rate and MSE. The momentum rate value is laying between zero and one and the commonly used range is around 0.9. At this figure MSE still zero up to the momentum rate value is equal to 0.8 at which MSE is about 1.8 and then saturated at this value. Fig. 8 shows the relation between momentum rate and number of iteration. At the momentum rate value of 0.8 the number of iteration is reached to 1000 and still saturated at this value.

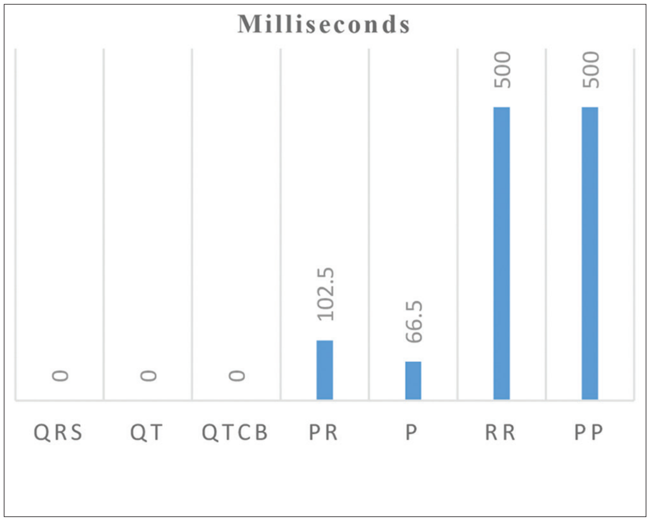


Fig. 14. Final diagnosis of electrocardiograph signals (sinus tachycardia case).

Fig. 9 demonstrates the relation between network size and network capacity. At this figure it is clear that there is a linear relation between network size and network capacity. As the network size increases up to 140,000 it is clear that the network capacity increases up to 35000. Fig. 10 shows the relation between network size and generalization. At this figure it is clear that the maximum generalization is obtained at starting point of the network size. Then the generalization decreases when the network size increases. On the other hand, the zero generalization is obtained at the network size equal to 1000.

Fig. 11 presents the normal case of ECG signal. Fig. 12 shows a normal patient having sinus normal rhythm in which the measurement ECG parameters are: QRS: 189.00 ms, QT: 292.66 ms, QTcB: 66.90 ms, PR: 137.10 ms, P: 88.85 ms, RR: 764.46 ms, and PP: 775.50 ms, this case indicated that the patient diagnosis is normal.

Fig. 13 deals with the sinus tachycardia case of ECG signal. Fig. 14 shows another patient having sinus tachycardia rhythm in which the measurement ECG parameters are: QRS: 0.0 ms, QT: 0.0 ms, QTcB: 0.0 ms, PR: 102.5 ms, P: 66.5 ms, RR: 500 ms, and PP: 500 ms, this case indicated that the patient diagnosis is sinus tachycardia.

6. CONCLUSIONS

The diagnosis of heart diseases depends largely on ECG, in addition to other devices that give special properties and parameters that leading to great importance in the field of healthcare. Measurements of ECG signals lead to the identification of problems experienced by people with heart disease. Real-time ECG diagnosis has several advantages as it is important in sharing private information in healthcare systems especially for heart diseases. The implemented approach accompanies the properties of features extracted from the lookup table and properties of neural networks. The feature extraction step verifies the features from the received ECG and neural networks give good responses to the new input patterns. The applied approach gives accurate detection of ECG signals as well as good quality of recognized ECG signals.

REFERENCES

[1] T. Gandhi, B. K. Panigrahi, M. Bhatia and S. Anand. (2010) “Expert model for detection of epileptic activity in EEG signature”. *Expert Systems with Applications,* vol. 37, pp. 3513-3520, 2010.

[2] S. Sanei and J. A. Chambers. “EEG Signal Processing”. John Wiley & Sons Ltd., Chichester, 2013.

[3] K. Polat and S. Günes. “Classification of epileptic form EEG using a hybrid system based on decision treeclassifier and fast Fourier transform”. *Applied Mathematics and Computation*, vol. 187, pp. 1017-1026, 2007.

[4] G. Ouyang, X. Li, C. Dang and D. A. Richards. “Using recurrence plot for determinism analysis of EEG recordings in genetic absence epilepsy rats”. *Clinical Neurophysiology*, vol. 119, pp. 1747-1755, 2008.

[5] M. Ahmadlou, H. Adeli and A. Adeli. “New diagnostic EEG markers of the Alzheimer’s disease using visibility graph”. *Journal of Neural Transmission*, vol. 117, no. 9, pp. 1099-1109, 2010.

[6] N. Kannathal, U. R. Acharya, C. M. Lim, Q. Weiming, M. Hidayat and P. K. Sadasivan. “Characterization of EEG: A comparative study”. *Computer Methods and Programs in Biomedicine*, vol. 80, no. 1, pp. 17-23, 2005.

[7] N. W. Willingenburg, A. Daffertshofer, I. Kingma and J. H. Van Dieen. “Removing ECG contamination from EMG recordings: A comparison of ICA-based and other filtering procedures”. *Journal of Electromyography and Kinesiology*, vol. 22, no. 3, pp. 485:493, 2010.

[8] C. Marque, C. Bisch, R. Dantas, S. Elayoubi, V. Brosse and C. Perot. “Adaptive filtering for ECG rejection from surface EMG recordings”. *Journal of Electromyography and Kinesiology*, vol. 15, no. 3, pp. 310-315, 2005.

[9] S. Abbaspour, M. Linden and H. Gholamhosseini. “ECG artifact removal from surface EMG signal using an automated method based on wavelet-ICA”. *Studies in Health Technology and Informatics*, vol. 211(pHealth), pp. 91-97, 2015.

[10] A. L. Hoff. “A simple method to remove ECG artifacts from trunk muscle EMG signals”. *Journal of Electromyography and Kinesiology*, vol. 19, no. 6, pp. 554-555, 2009.

[11] P. E. McSharry, G. Clifford, L. Tarassenko and L. A. Smith. “A dynamical model for generating synthetic electrocardiogram signals”. *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 289-294, 2003.

[12] M. S. AL-Ani and A. A. Rawi. “ECG beat diagnosis approach for ECG printout based on expert system”. *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 4, pp. 797-807, 2013.

[13] M. S. AL-Ani and A. A. Rawi. “Rule-based expert system for automated ECG diagnosis. *International Journal of Advances in Engineering and Technology,* vol. 6, no. 4, pp. 1480-1493, 2013.

[14] J. E. Madias, R. Bazaz, H. Agarwal, M. Win and L. Medepalli. “Anasarca-mediated attenuation of the amplitude of electrocardiogram complexes: A description of a heretofore unrecognized phenomenon”. *Journal of the American College of Cardiology*, vol. 38, no. 3, pp. 756-764, 2001.

[15] U. R. Acharya, V. K. Sudarshan, H. Adeli, J. Santhosh, J. E. W. Koh, S. D. Puthankatti and A. Adeli A. “A novel depression diagnosis index using nonlinear features in EEG signals”. *European Neurology*, vol. 74, no. 79-83, 2015.

[16] K. N. Khan, K. M. Goode, J. G. F. Cleland, A. S. Rigby, N. Freemantle, J. Eastaugh, A. L. Clark, R. de Silva, M. J. Calvert, K. Swedberg, M. Komajda, V. Mareev, F. Follath and EuroHeart Failure Survey Investigators. “Prevalence of ECG abnormalities in an international survey of patients with suspected or confirmed heart failure at death or discharge. *European Journal of Heart Failure*, vol. 9, pp. 491-501, 2007.

[17] K. Y. K. Liao, C. C. Chiu and S. J. Yeh. “A novel approach for classification of congestive heart failure using relatively short-term ECG waveforms and SVM classifier. In: *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, IMECS March 2015, Hong Kong, pp. 47-50, 2015.

[18] R. J. Martis, U. R. Acharya and C. M. Lim. “ECG beat classification using PCA, LDA, ICA and discrete wavelet transform”. *Biomedical Signal Processing and Control,* vol. 8, no. 5, pp. 437-448, 2013.

[19] U. Orhan. “Real-time CHF detection from ECG signals using a novel discretization method”. *Computers in Biology and Medicine*, vol. 43, pp. 1556-1562, 2013.

[20] J. Pan and W. J. Tompkins. “A real time QRS detection algorithm”. *IEEE Transactions on Biomedical Engineering,* vol. 32, no. 3, 1985.

[21] M. Sadaka, A. Aboelela, S. Arab and M. Nawar. Electrocardiogram as prognostic and diagnostic parameter in follow up of patients with heart failure. *Alexandria Journal of Medicine*, vol. 49, pp. 145-152, 2013.

[22] K. Senen, H. Turhan, A. R. Erbay, N. Basar, A. S. Yasar, O. Sahin and E. Yetkin. “P wave duration and P wave dispersion in patients with dilated cardiomyopathy”. *European Journal of Heart Failure*, vol. 6, pp. 567-569, 2004.

[23] R. A. Thuraisingham. “A classification system to detect congestive heart failure using second-order difference plot of RR intervals”. *Cardiology Research and Practice*, vol. 2009, p. ID807379, 2009.

[24] E. D. Ubeyli. “Feature extraction for analysis of ECG signals”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Milano, Italy, pp. 1080-1083, 2008.

[25] R. Rodríguez, A. Mexicano, J. Bila, S. Cervantes and R. Ponce. “Feature extraction of electrocardiogram signals by applying adaptive threshold and principal component analysis”. *Journal of Applied Research and Technology*, vol. 13, pp. 261-269, 2015.

[26] H. Gothwal, S. Kedawat and R. Kumar. “Cardiac arrhythmias detection in an ECG beat signal using fast Fourier transform and artificial neural network”. *Journal of Biomedical Science and Engineering*, vol. 4, pp. 289-296, 2011.

[27] S. A. Chouakri, F. Bereksi-Reguig, A. T. Ahmed. “QRS complex detection based on multi Wavelet packet decomposition”. *Applied Mathematics and Computation*, vol. 217, pp. 9508-9525, 2011.

[28] D. S. Benitez, P. A. Gaydecki, A. Zaidi and A. P. Fitzpatrick. “A new QRS detection algorithm based on the Hilbert Transform”. *Computers in Cardiology*, vol. 2000, pp. 379-382, 2000.

[29] G. Vijaya, V. Kumar and H. K. Verma. “ANN-based QRS-complex analysis of ECG”. *Journal of Medical Engineering and Technology*, vol. 22, pp. 160-167, 1998.

[30] M. Ayat, M. B. Shamsollahi, B. Mozaffari and S. Kharabian. “ECG denoising using modulus maxima of wavelet transform”. In: *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*: Engineering the Future of Biomedicine EMBC, pp. 416-419, 2009.

[31] F. Chiarugi, V. Sakkalis, D. Emmanouilidou, T. Krontiris, M. Varanini and I. Tollis. “Adaptive threshold QRS detector with best channel selection based on a noise rating system”. *Computers in Cardiology*, vol. 2007, pp. 157-160, 2007.

[32] M. Elgendi. “Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases”. PLoS One, vol. 8, p. e73557, 2013.

[33] A. Rehman, M. Mustafa, I. Israr and M. Yaqoob. “Survey of wearable sensors with comparative study of noise reduction ECG filters”. *International Journal of Computing and Network Technology*, vol. 1, pp. 61-81, 2013.

[34] M. Elgendi, B. Eskofier and D. Abbott. “Fast T wave detection calibrated by clinical knowledge with annotation of P and T waves”. *Sensors (Basel*), vol. 15, pp. 17693-17714, 2015.

[35] M. Rahimpour, B. M. Asl. “P wave detection in ECG signals using an extended Kalman filter: An evaluation in different arrhythmia contexts”. *Physiological Measurement*, vol. 37, pp. 1089-1104, 2016.

[36] A. Z. Mohammed, A. F. Al-Ajlouni, M. A. Sabah and R. J. Schilling. “A new algorithm for the compression of ECG signals based on mother wavelet parameterization and best-threshold levels selection”. *Digital Signal Processing*, vol. 23, pp. 1002-1011, 2013.

[37] B. M. Reza, R. Kaamran and K. Sridhar. “Robust ultra-low-power algorithm for normal and abnormal ECG signals based on compressed sensing theory”. *Procedia Computer Science*, vol. 19, pp. 206-213, 2013.

[38] M. E. Ann and M. A. Andrés. “Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals”. *Computational Statistics and Data Analysis*, vol. 70, pp. 67-87, 2014.

[39] M. H. Vafaie, M. Ataei and H. R. Koofigar. “Heart diseases prediction based on ECG signals’ classification using a genetic fuzzy system and dynamical model of ECG signals”. *Biomedical Signal Processing and Control*, vol. 14, pp. 291-296, 2014.

[40] A. Kamal and A. Nader. “Design and implementation of a multiband digital filter using FPGA to extract the ECG signal in the presence of different interference signals”. *Computers in Biology and Medicine*, vol. 62, pp. 1-13, 2015.

[41] E. Farideh, S. Seyed-Kamaledin and N. Homer. “Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs”. *Biomedical Signal Processing and Control,* vol. 18, pp. 69-79, 2015.

[42] S. H. Shirin and M. Behbood. “A new personalized ECG signal classification algorithm using Block-based Neural Network and Particle Swarm Optimization”. *Biomedical Signal Processing and Control*, vol. 25, pp. 12-23, 2016.

[43] Y. Om Prakash and R. Shashwati. “Smoothening and Segmentation of ECG signals using total variation denoising, minimization, majorization and bottom-up approach”. *Procedia Computer Science*, vol. 85, pp. 483-489, 2016.

[44] M. Aleksandar and G. Marjan. “Improve d pipeline d wavelet

implementation for filtering ECG signals". *Pattern Recognition Letters*, vol. 95, pp. 85-90, 2017.

[45] M. Kumar, R. B. Pachori and U. R. Acharya. "Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals". *Biomedical Signal Processing and Control*, vol. 31, pp. 301-308, 2017.

[46] M. S. Al-Ani. "Electrocardiogram waveform classification based on P-QRS-T Wave recognition." *UHD Journal of Science and Technology*, vol. 2, no. 2, pp. 7-14, 2018.

[47] J. J. Nallikuzhy and S. Dandapat. "Spatial enhancement of ECG using multiple joint dictionary learning". *Biomedical Signal Processing and Control*, vol. 54, p. 101598, 2019.

[48] C. Han and L. Shi. "ML–ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG." *Computer Methods and Programs in Biomedicine*, vol. 185, p. 105138, 2020.

[49] L. A. Abdulla and M. S. Al-Ani. "A review study for electrocardiogram signal classification". *UHD Journal of Science and Technology (UHDJST)*, vol. 4, no. 1, 2020.

[50] L. A. Abdullah and M. S. Al-Ani. "CNN-LSTM based model for ECG arrhythmias and myocardial infarction classification". *Advances in Science Technology and Engineering Systems Journal*, vol. 5, no. 5, pp. 601-606, 2020.

REVIEW ARTICLE

UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Construction of Alphabetic Character Recognition Systems: A Review

Hamsa D. Majeed*, Goran Saman Nariman

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

ABSTRACT

Character recognition (CR) systems were attracted by a massive number of authors' interest in this field, and lot of research has been proposed, developed, and published in this regard with different algorithms and techniques due to the great interest and demand of raising the accuracy of the recognition rate and the reliability of the presented system. This work is proposed to provide a guideline for CR system construction to afford a clear view to the authors on building their systems. All the required phases and steps have been listed and clarified within sections and subsections along with detailed graphs and tables beside the possibilities of techniques and algorithms that might be used, developed, or merged to create a high-performance recognition system. This guideline also could be useful for readers interested in this field by helping them extract the information from such papers easily and efficiently to reach the main structure along with the differences between the systems. In addition, this work recommends to researchers in this field to comprehend a specified categorical table in their work to provide readers with the main structure of their work that shows the proposed system's structural layout and enables them to easily find the information and interests.

Index Terms: Optical Character Recognition, Script Identification, Document Analysis, Character Recognition, Multi-Script Documents

1. INTRODUCTION

In recent decades, many studies have demonstrated the ability of the machine to examine the environment and learn to distinguish patterns of interest from their background and make reliable and feasible decisions regarding the categories of the patterns. With huge volumes of data to be dealt with and through years of research, the design of approaches based on character recognition (CR) remains an ambiguous goal. Various frameworks employed machine learning approaches which have been most comprehensively studied and applied to a large number of systems that are essential in building a high-accuracy recognition system, CR is among the most well-known techniques and methods that make use of such artificial intelligence which have received attention increasingly. Moreover, in various application domains, ranging from computer vision to cybersecurity, character classifiers have shown splendid performance [1]-[3].

The application of CR is concerned with several fields of research. Through those numerous applications, there is no single approach for recognition or classification that is optimal and that motivates the researchers to explore multiple methods and approaches to employ. In addition, a combination of several techniques and classifiers is popped to the surface to serve the same purpose. Due to the increased attention paid to CR-based applications, noticeably there are few comprehensive overviews and systematic mappings of

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp32-42	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: 09-11-2022 Accepted: 07-02-2023 Published: 18-02-2023

CR applications design. Instead, the existing reviews explore in detail a specific domain, technique, or system focusing on the algorithms and methodology details [4], [5].

While starting investigations in this field, a big space of confusion appeared while diving into the details of each step in the recognition process due to the variety of paths that could be taken to reach the final goal and the pool of factors to be phished for that matter. That leads to the fact of considering an in-depth literature review as a requirement for surveying the possibility of using the techniques, approaches, or methodologies that are required for that phase of the recognition process among the others and deciding if they are suitable or not for that CR-based application.

The major aim of this study is to present the main path for the various kinds of approaches to be followed before diving into the details of the framework to be proposed by the meant research, Moreover, depending on each research field, there are options offered and categorized, techniques, and methods are presented and summarized from multiple perspectives all of which are investigated to answer the following queries:

1. Which language will be taken to recognize as input and what is a specified script writing style?
2. How can the data be acquired? Is it taken digitally (touch-screen, scanner, or another digital device) or uploaded from a non-digital source? In printed form by a keyboard or in handwritten form?
3. Which scale or level of detail is present in that set of data? Does the script have to be taken wholly or by a single character each time?
4. From which source could those data be collected? Is the preprocessing phase needed or not?
5. Generally, through which recognition process should invade for the optimal outcomes considering the previously chosen phases?

This work is structured to give the most suitable roadmap to the author of interest by presenting a systematic guideline to explore the multidisciplinary path starting from the script writing style the passing by the most suitable guide throughout the desired dataset characteristics (acquisition, granularity level, and the source of collected data), reaching to the script recognition process for the CR-based applications. Furthermore, this study uncovers the potential of CR applications among different domains and specifications by summarizing the purpose, methodologies, and application.

Thorough proofreading of several types of research including survey articles, the CR process has the same stations to stop by which could be sorted under some separated categories on specific factors and all those categories of any proposed system may have a stop in those main stations, that was an encouragement to make this study to highlight those main stations and present a guideline the researchers of interest by examining the detailed of sub-stations due to building CR system efficient to the author and understandable by the reader.

2. PROPOSED WALKTHROUGH GUIDELINE

The main goal of this study is to construct and design criteria for researchers working in the field of CR systems to observe when initiating research in both the practical and written parts. The following classifications and assortments are proposed, as shown in Fig. 1.

2.1. Script Writing System

From the linguistic point of view, nowadays, scripts used throughout the work have been broken down into six script classes, each of which can be used in one or more languages [6], [7]. Furthermore, in the context of CR, the investigations of the script character characteristics and structural properties, the script-written system has been classified under six classes. Different classes may contain the same language scripts [8], [9], [10]. Fig. 2 illustrates the classification of the script writing system.

2.1.1. Logographic system

The oldest kind of writing system is a logographic writing system; it is also called an ideogram as well, which employs symbols to depict a whole word or morpheme. The most well-known logographic script is Chinese, but logograms

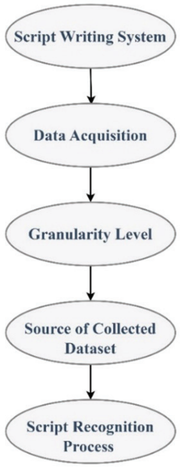


Fig. 1. General assortments of the CR system.

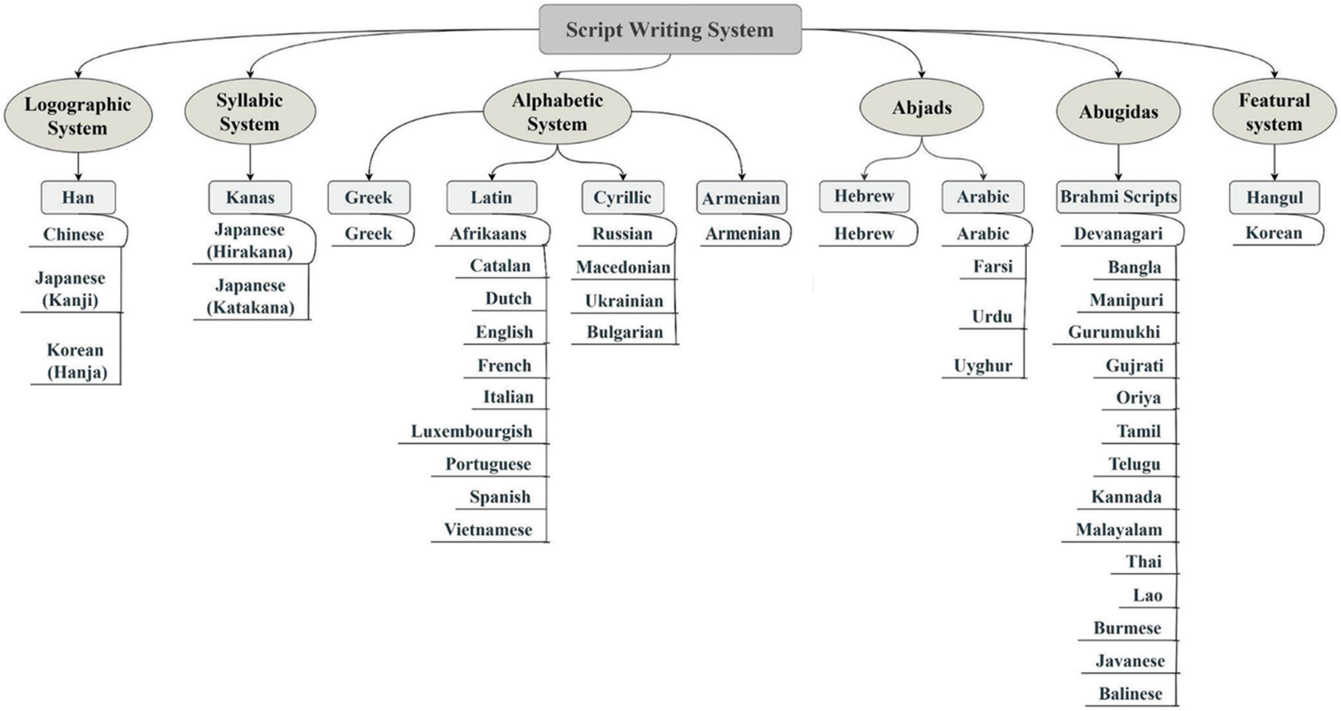


Fig. 2. Script Written System classifications.

such as numbers and the ampersand are found in almost all languages. An ideographic writing system typically has thousands of characters. Thus, the recognition process of this kind of script is still a challenging and fascinating topic for researchers. Han is the only script family in this class that includes two more languages, namely, Japanese (Kanji) and Korean (Hanja). The interesting distinguishing point between Han and other languages is the text line written direction, which is either from top to bottom or left to right.

In literature, lots of research can be found on handwritten CR in these scripts, for instance [11]-[13] work on Chinese, Japanese (Kanji), and Korean (Hanja), respectively. The accuracy rates for the scripts based on the aforementioned references are 99.39%, 99.64%, and 86.9%, respectively.

2.1.2. Syllabic system

Every written sign in a syllabic system, such as the one used in Japanese, corresponds to a phonetic sound or syllable. Kanas, which are divided into two types - Hirakana and Katakana - represent Japanese syllables. The Japanese script combines logographic Kanji and syllabic Kanas, as mentioned in the previous subsection. The Kanas has a similar visual appearance to the Chinese, with the exception that the Kanas has a lower density than the Chinese.

A lot of recognition progress can be found in the literature for both Hirakana and Katakana. Examples of excellent achievements in recognition accuracy rate are contributed in [11] for both Hirakana and Katakana, which are 98.83% and 98.19%, respectively.

2.1.3. Alphabetic system

Each consonant and vowel have a distinct symbol in the alphabetic writing system, which is used to write the languages classified under this written system. Segmental systems are another name for alphabets. To represent spoken language, these systems mix a small number of characters called letters. Letters are meant to represent certain phonemes. Greece is where the alphabet was first used, and it later expanded around the world, especially in Europe and a part of Asia as well [14], proposed a system for Ancient Greek CR that achieved an accuracy rate of 96%. Latin, Cyrillic, and Armenian also belong to this system.

There are numerous languages that use the Latin alphabet, commonly known as the Roman script, with differing degrees of alteration. It is utilized to write in a wide range of European languages, including English, French, Italian, Portuguese, Spanish, German, and others. The interested authors of Latin languages presented their ideas in terms of the recognition system for the different Latin languages, for instance, Afrikaans 98.53% [15], Catalan 91.97% [16], Dutch 95.5% [17], English 98.4% [18], French 93.6% [19], Italian 92.47% [20], Luxembourgish (87.55 ± 0.24)% [21], Portuguese 93% [22], Spanish 97.08% [23], Vietnamese 97% [24], and German 99.7% [25].

Cyrillic has a separate letter set but is still relatively comparable to Latin. The Cyrillic writing system has been adopted by certain Asian and Eastern European languages, including Russian, Bulgarian, Ukrainian, and Macedonian, where the recognition rate is recorded for them as follows: Russian 83.42% [26], Bulgarian 89.4% [27], Ukrainian 95% [28], and Macedonian 93% [29].

Finally, the Armenian written system, this language classified as an Indo-European language belonging to an independent branch of which it is the only member recent CR system for this language scored 89.95% [30].

2.1.4. Abjads

When the words have a writing pattern from right to left along with text line, written in a repetition of consonants that are close together leaving the vowel sounds to be inferred by the reader, and have cursive long strokes consisting of few dots, then you are looking at Abjads writing system. It is unlike most other scripts in the world but it is similar to the alphabetic system unless it has symbols for consonantal sounds only. These unique features make the process of script identification for Abjads relatively simpler compared to other scripts, particularly because of the long cursive strokes with dots and the right-to-left writing direction, making it easier for recognition systems in pen computing.

Arabic and Hebrew are considered the major categories of the Abjads writing system. There are some other scripts of Arabic origin, such as Farsi (Persian), Urdu, and Uyghur. A lot of approaches had been proposed for identifying Abjad-based scripts, they used the long main stroke along with the cursive appearance yielding from conjoined words for Arabic. Meanwhile, the more uniform strokes in length and discrete letters were the main dependent features of Hebrew script recognition. According to the latest survey for Arabic recognition systems [31], the highest accuracy score is 99.98%, while recorded 97.15% for Hebrew [32]. In Farsi, Urdu, and Uyghur, the highest accuracies achieved are 99.45%, 98.82%, and 93.94%, respectively [33]-[35].

2.1.5. Abugidas

It is a writing script primarily based on a consonant letter and secondary vowel notation. They are sharing with alphabetic systems the property of combining characters writing styles within the text line. It belongs to the Brahmic family of scripts which is can be expressed in two groups:

1. Original Brahmi script: This northern group deployed in Devnagari, Bangla (Bengali), Manipuri, Gurumukhi, Gujrati, and Oriya languages. The most recent survey papers for the CR systems of this group come up with the highest recognition rate of 99% for Devnagari, 99.32% for Bangla (Bengali), 98.70% for Manipuri, 99.3% for Gurumukhi, 98.78% for Gujrati, and 96.7% for Oriya [36]-[38].
2. Derived from Brahmi: Look quite different from the northern group and used in:
   a. South India: Tamil, Telugu, Kannada, and Malayalam, where the highest accuracy of the mentioned language for recognition matter was for Tamil 98.5%, Telugu 98.6%, Kannada 92.6%, and Malayalam 98.1% [39].
   b. Southeast Asia: Thai, Lao, Burmese, Javanese, and Balinese, the languages of this group have achieved the highest validation rate where Thai, Lao, and Burmese attained 92.1%, 92.41%, and 96.4% while Javanese and Balinese gained 97.7% and 97.53%, respectively [40]-[43].

2.1.6. Featural system

This form of writing system is significantly represented by symbols or characters, the main language is Korean which is described as less complex and less dense compared to Chinese and Japanese, it is represented by mixing logographic Hanja and featural Hangul, the highest scored accuracy rate for Korean was 97.07% [44].

As a summarization of all the findings in this section, Table 1 illustrates the classifications of the languages with the highest accuracy recorded so far.

2.2. Data Acquisition

The next step for the author after selecting which language to work on is to decide which writing style will be chosen for recognition, this step is considered one of the fixed and essential phases in all the recognition studies and research, reaching this phase requires the knowledge of how to start acquiring data to be fed into the recognition system, the answer simply starts with defining the writing style, here the author has two options either printed script or handwritten script.

After making the decision, the acquisition tools are required either offline tools or online. In this section, a guideline is

TABLE 1: Summarization of languages with their recent highest accuracy rate

Script writing system	Main language	Sub-language	Accuracy rate (%)
Logographic system	Han	Chinese	99.39
		Japanese (Kanji)	99.64
		Korean (Hanja)	86.9
Syllabic system	Kanas	Japanese (Hirakana) a	98.83
		Japanese (Katakana)	98.19
Alphabetic system	Greek	Greek	96
	Latin	Afrikaans	98.53
		Catalan	91.97
		Dutch	95.5
		English	98.4
		French	93.6
		Italian	92.47
		Luxembourgish	(87.55±0.24)
		Portuguese	83
		Spanish	97.08
		Vietnamese	97
		German	99.7
	Cyrillic	Russian	83.42
		Bulgarian	89.4
		Ukrainian	95
		Macedonian	93
	Armenian	Armenian	89.95
Abjads	Hebrew	Hebrew	97.15
	Arabic	Arabic	99.98
		Farsi	99.45
		Urdu	98.82
		Uighur	93.94
Abugidas	Brahmi	Devnagari	99
		Bangla (Bengali)	99.32
		Manipuri	98.70
		Gurumukhi	99.3
		Gujrati	98.78
		Oriya	96.7
		Tamil	98.5
		Telugu	98.6
		Kannada	92.6
		Malayalam	98.1
		Thai	92.1
		Lao	92.41
		Burmese	96.4
		Javanese	97.7
		Balinese	97.53
Featural system	Korean	Korean	97.07

proposed and could be followed to help make those decisions as Fig. 3 shows.

2.2.1. Printed character

Those characters are produced as a result of the process of producing using inked-type tools. In recognition systems of any language, the printed characters usually achieve a high recognition rate because it is considered in regular form, clean, have the same style, and have similar shapes and lines, and that facilitates the learning operation and therefore raises the accuracy of recognition in the testing phase.

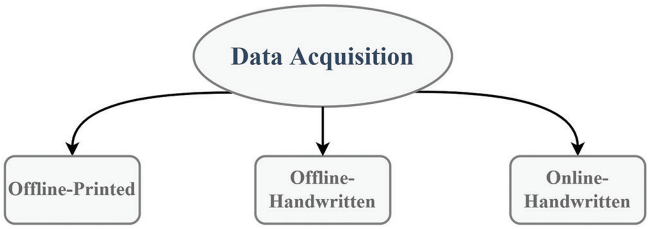


Fig. 3. Overview of data acquisition.

2.2.2. Handwriting character

When the process of forming letters of any language is done with the hand, rather than any typing device then the result is handwriting characters. Most of the authors that are interested in CR are employing handwriting characters as input to their approaches to prove the effectiveness and efficiency of their systems or techniques due to the complexity and impenetrability that come with the variety of the handwriting style and the use of tools besides the differences in lines and colors not to mention the irregular shapes and positions.

2.2.3. Online character

These characters are obtained from digital devices with a touch screen with/without a keyboard involved like a personal digital assistant, or mobile. Where screen sensors receive the switching of pushing and releasing the pen on the screen in addition to the pen tip movements over the screen.

2.2.4. Offline character

This kind of character is attended when 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.

It is essential for the author to choose the correct combination of the writing style and the writing tool, as Fig. 3 illustrates there are three combinations to decide among them: offline-printed where the input of the CR system decided to be in offline mode with characters taken from the printed device rather than the offline-handwritten which taken from a human-hand in offline-mode already written on paper in a previous time while the online-handwritten fed as input to CR system instantly by hand through a touchable input device without a keyboard.

Some recent recognition systems are illustrated in Table 2 for several languages to show some authors' choices for the language, writing style, and writing tool, and how their choices affect the accuracy rate for each mechanism. Furthermore, a comprehensive survey for online and offline handwriting recognition can be found in Plamondon and Srihari [45].

The outcomes of Table 2 show that most existing studies have focused on handwritten text, with fewer works attempting to classify or identify printed text. This is because of the high variance in handwriting styles across people and the poor quality of the handwritten text compared to printed text yields the fact that handwritten CR is more challenging than the printed one.

On the other hand, it is noticeable using offline as writing tool more than online ones this is due to in the online case, features can be extracted from both the pen trajectory and the resulting image, whereas in the offline case, only the image is available, so the offline recognition is observed as harder than online recognition.

2.3. Granularity Level of Documents

The third type of classification of character handwriting recognition is "Granularity Level of Documents," which describes the level of detailed information taken as initial input to the defined and proposed framework. This class could be split into five granularity levels as shown in Fig. 4, from a script page full of text to a single letter or symbol.

TABLE 2: Examples of recognition systems with different data acquisition mechanisms

Reference	Language	Writing style	Writing tool	Accuracy rate (%)
[46]	Arabic	Handwritten	Offline	99.93
[18]	English	Handwritten	Offline	98.4
[47]	English	Printed	Offline	98
[48]	English	handwritten	Online	93.0
[49]	Chinese	Handwritten	Online	98
[50]	Chinese	Handwritten	Offline	94.9
[13]	Chinese	Printed	Offline	99.39
[51]	Arabic	Printed	Offline	97.51
[52]	Arabic	Handwritten	Online	96

In the domain of CR, if the initial input into the OCR framework is not at a character level, the process of script identification must proceed until it gets to a single character. This procedure, known as "Segmentation," will be covered in the following subsection (3.5).

2.3.1. Document/page level

Document-level script is the most detailed granularity level, where the entire document is exposed to the script identification procedure at once. Following processing, the document is further broken down into pages, pieces of paragraphs, text lines, words, and finally characters to enable the recognition of the precise letter. Although some researchers discriminate between the script recognition process at the document and page levels, in general, the technical methodologies are very similar. Because of this, some researchers alternately refer to document-level and page-level script recognition.

Finding the text region on a page is the initial step in page-level script identification. It is possible to carry out this operation by separating the pages into text and non-text pieces [53]. Several pieces of research can be found in the literature for both offline-handwritten [54] and offline-printed [55].

After the page of the script has been identified, the process of the next level starts, which is paragraph or text block identification. It operates by dividing the entire page into equal-sized text blocks with several lines of content. Text blocks can have different sizes, and padding may be necessary if characters are on the edge of a text block [56]. Is an example of segmenting pages into pieces of text blocks.

2.3.2. Paragraph level

The text block is separated into lines. The white space between lines is typically used for text line segmentation. Lines of scrip are detected and segmented to be prepared for further segmentation processing. Both offline-handwritten [57] and offline-printed [58] line detection has been the subject of numerous studies in the literature.

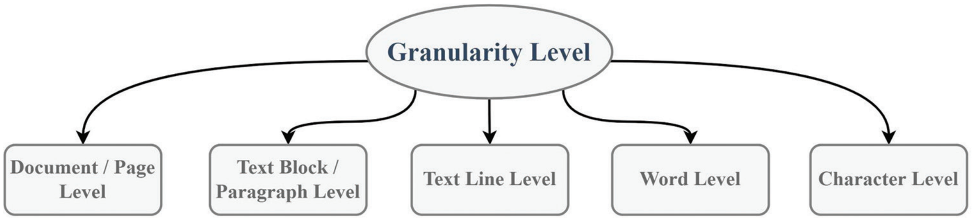


Fig. 4. Granularity level classification.

2.3.3. Text line level

A framework that gets a line of script as initial input needs segmentation processes to identify each word in the text. Therefore, word identification is needed. Text lines are divided into words; usually, the white space between text lines is used for this purpose. Numerous literary attempts have been made to address the difficulties encountered in this process. For instance, there might be noise, twisted words, missing or partial letters in words, words that are not available as straight text lines, etc. Some examples given in this topic of identifying words in a text line are [59] for offline-handwritten and [60] for offline-printed.

2.3.4. Word level

Character detection and segmentation are required since the initial input is a word. It usually works by combining properties from various characters to ensure the process. Several attempts have been made to improve accuracy and ensure that no character inside a word is missed. For instance, recently [61] used distinct strategies and achieved a satisfactory outcome.

2.3.5. Character level

Finally, there is no requirement for segmentation at the character level because the initial input into the proposed framework is already character. The character goes through preprocessing, which is followed by recognition procedures. In some circumstances, no preprocessing is required, as is the case when using a character public dataset. For instance [62], is an example of working at the character level with and without preprocessing, respectively.

In addition, to avoid confusion between granularity levels for identification/detection and recognition processes, it is worth mentioning that from the recognition standpoint, when the granularity level is text line level, it means that the text line is already known and the detection and segmentation into words and characters are needed. However, from the identification/detection point of view, it means that the identification and detection of text lines are working. Further details about these processes can be found in [10], [63].

2.4. Source of Collected Dataset

The essential component of any machine learning application is the dataset. That leads us to discuss this important phase of CR as the fourth classification named Source of Collected Dataset which is broken down into two categories as Fig. 5 illustrates:

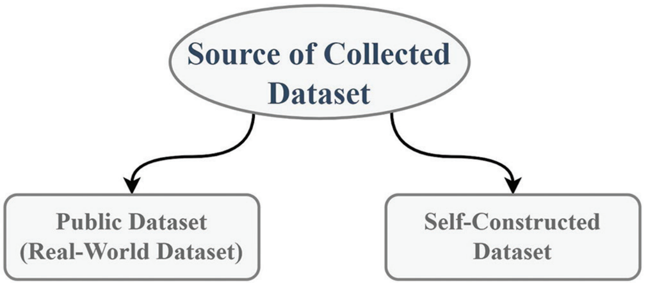

Source of Collected Dataset
Public Dataset (Real-World Dataset)
Self-Constructed Dataset

Fig. 5. Categories of collected dataset sources.

2.4.1. Public dataset (real-world dataset)

The term "public dataset" refers to a dataset saved in the cloud and made open to the public. MNIST, Keras, Kaggle, and others are examples. Almost all of the public datasets have been preprocessed, cleaned, and usually, in the case of character level, reshaped to 28 × 28 pixels and saved as CSV files. Many authors attain to use this source to skip the preprocessing step and focus more on the other steps and easily find opponents for the comparison issue of those who used the same data source with different techniques.

2.4.2. Self-constructed dataset

Is the dataset that the researchers create and prepare on their own depending on their techniques, it is an online or offline way of collection, this source of dataset is considered more challenging because the collected images are not processed at all in terms of resizing, denoising, colored, etc.

For a fair comparison, this kind of work better to be compared with studies that have done with a self-collected source of data, not with a public one that comes clean and processed. Researchers should be aware of the data to be collected and use the proper tools required to preprocess in a way that suits the technique used for recognition.

2.5. Script Recognition Process

The script recognition process (the implementable phase) is the fifth classification type of alphabet handwritten recognition framework. In an in-depth study of several research articles, including survey articles, we mainly focused on the phases that an OCR system needs to accomplish its recognition goal. Thus, we could conclude that four categories can be defined based on the number of phases in which the whole procedure of recognition comprises, as presented in Fig. 6.

In addition, commonly, script recognition is achieved by blending traditional image processing techniques with

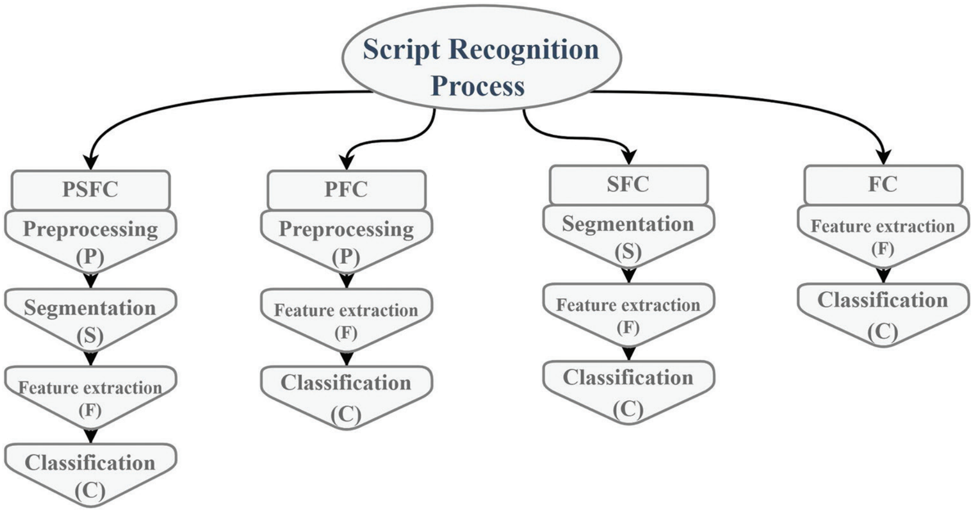


Fig. 6. Basic components of the script recognition processes.

image identification and recognition techniques. The recognition composition is formed from four primary phases, namely, Preprocessing (P), Segmentation (S), Feature Extraction (F), and Classification (C). The last two phases, Feature-Extraction and Classification, are the most common in the research. There is not any work without any of these two phases. The next few paragraphs will briefly outline them.

- Preprocessing (P) is a sequence of operations performed to intensify the input image. It is responsible for removing noise, resizing, thinning, contouring, transforming the dataset into a black-white format, edge detection, etc. Every single one of them can be performed with an appropriate technique
- Segmentation (S) performs the duty of obtaining a single character. The document processing follows a hierarchy; it starts from the whole page and ends with a single character. The required level of the hierarchy is a single character
- Feature extraction (F) is a mechanism in which each character is turned into a feature vector using specific algorithms for the extraction of the features, which is then fed into a classifier to determine which class it belongs to.
- The classification (C) phase is a decision-making process that uses the features extracted from the preceding step as input. And it decides what the final output is.

It is worth noticing that handwritten mathematical symbols and expressions recognition is out of our research scope. Therefore, we do not consider the two additional phases (Structural Analysis and Symbol Recognition) which are included in such works. More details can be found in Sakshi and Kukreja [64].

2.5.1. PSFC

The first category of the Script Recognition Process class can be called PSFC, which means all four phases have been utilized to achieve the goal as [65] describe.

2.5.2. PFC

The segmentation process is skipped in the second category, in most cases due to working on character level as initial input therefore no need for segmentation as presented in Parthiban *et al.* [66].

2.5.3. SFC

The third one is SFC as [67] proposal, where the preprocessing is missed because the entered data originally is clean and there is no preprocessing required.

2.5.4. FC

In the fourth and last category as illustrated in Gautam and Chai [68], the first two phases P and F are dismissed because the granularity level is letters, and the initial input data is originally clean. For instance, works utilizing public datasets such as MNIST [69] could be classified under this category.

3. EXAMPLES

This section is to illustrate some of the CR systems and gives a description of how to read their roadmap regarding their systems, by applying the proposed guideline, any paper in this field can be summarized in stages according to the

TABLE 3: Examples of the proposed framework of character recognition

Example 4 [48]		Example 5 [13]	
Classifications	**Nominated category**	**Classifications**	**Nominated category**
Script writing system	English	Script writing system	Chinese
Data acquisition	Online-handwritten	Data acquisition	Offline-printed
Granularity level of documents	Line level	Granularity level of documents	Character level
Source of the collected dataset	Public dataset	Source of the collected dataset	Self-constructed dataset
Script recognition process	FC	Script recognition process	PFC

author's choices and be easier to the reader to figure out the main stages and for the other authors to develop any desired CR system.

Some examples are presented here to show how the CR system can be summarized according to the proposed guideline, and resembling a table is suggested to be created in such a work to provide a comprehensive view of the proposed framework as a whole. It makes it easier for the reader to find the information they are searching for before going into depth. Table 3 provides two examples of how to present the suggested table. In addition, the following examples demonstrate how the systems may be constructed using the component chain:

1. [18] English → offline-handwritten → character level → self-constructed dataset → PFC
2. [46] Arabic → offline-handwritten → line level → public dataset → PSFC
3. [49] Chinese → online-handwritten → page level → public dataset → SFC
4. [48] English → online-handwritten → line level → public dataset → FC
5. [13] Chinese → offline-printed → character level → self-constructed dataset → PFC

4. CONCLUSION

CR stepped ahead as an eminent topic of research. Exhaustive studies continuously presented CR of different languages with various algorithms that were developed to increase the reliability of these characters for accurate recognition. A guideline for the construction CR system has been proposed for the authors in this field to overcome the unclear presentation and expressing ideas in such a domain of science. Almost all the required steps have been shown and demonstrated by graph and table to be used in such works in CR Domaine for more clarity for the authors to margin their scope. It is also for the readers, as well, to directly recognize the used technique through in-text reading and then move forward to the details afterward. Through reading this guideline, the authors will be able to order their thoughts and build their recognition system smoothly and effectively especially for the new authors in this field, as for readers after reading this work they will have the ability to analyze other research in the relative fields and extract information easily from other works of interest, for the seekers of new ideas or merging techniques, this guideline is suitable to help to determine the exact part of recognition system to be studied or compared with. Saving time, effort, and thoughts orienting for other authors or readers was one of the essential aims of this work.

REFERENCES

[1] M. Paolanti and E. Frontoni. "Multidisciplinary pattern recognition applications: A review". *Computer Science Review*, vol. 37, pp. 100276, 2020.

[2] M. Kawaguchi, K. Tanabe, K. Yamada, T. Sawa, S. Hasegawa, M. Hayashi and Y. Nakatani. "Determination of the Dzyaloshinskii-Moriya interaction using pattern recognition and machine learning". *npj Computational Materials*, vol. 7, no. 1, 2021.

[3] B. Biggio and F. Roli. "Wild patterns: Ten years after the rise of adversarial machine learning". *Pattern Recognition*, vol. 84, pp. 317-331, 2018.

[4] T. S. Gorripotu, S. Gopi, H. Samalla, A. V. Prasanna and B. Samira. "Applications of Computational Intelligence Techniques for Automatic Generation Control Problem-a Short Review from 2010 to 2018." In: Computational Intelligence in Pattern Recognition. Springer Singapore, Singapore, 2020, pp. 563-578.

[5] M. I. Sharif, J. P. Li, J. Naz and I. Rashid. "A comprehensive review on multi-organs tumor detection based on machine learning". *Pattern Recognition Letters*, vol. 131, pp. 30-37, 2020.

[6] A. Nakanishi. "Writing Systems of the World: Alphabets, Syllabaries, Pictograms". Charles E. Tuttle Co., United States, 1980.

[7] F. Coulmas. "The Blackwell Encyclopedia of Writing Systems". Blackwell, London, England, 1999.

[8] 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.

[9] D. Ghosh, T. Dube and A. P. Shivaprasad. "Script recognition--a review". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142-2161, 2010.

[10] 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.

[11] C. Tsai. "Recognizing handwritten Japanese Characters using Deep Convolutional Neural Networks". University of Stanford in Stanford, California, pp. 405-410, 2016.

[12] S. Purnamawati, D. Rachmawati, G. Lumanauw, R. F. Rahmat and R. Taqyuddin. "Korean letter handwritten recognition using deep convolutional neural network on android platform". *Journal of Physics Conference Series*, vol. 978, no. 1, p. 012112, 2018.

[13] Y. Q. Li, H. S. Chang and D. T. Lin. "Large-scale printed Chinese character recognition for ID cards using deep learning and few samples transfer learning". *Applied Sciences*, vol. 12, no. 2, p. 907, 2022.

[14] B. Robertson and F. Boschetti. "Large-scale optical character recognition of ancient Greek". *Mouseion Journal of the Classical Association of Canada*, vol. 14, no. 3, pp. 341-359, 2017.

[15] J. Hocking and M. Puttkammer. "Optical Character Recognition for South African languages". In: 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016.

[16] A. Fornes, V. Romero, A. Baró, J. I. Toledo, J. A. Sánchez, E. Vidal, J. Lladós. "ICDAR2017 Competition on Information Extraction in Historical Handwritten Records". In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017.

[17] H. van Halteren and N. Speerstra. "Gender recognition on Dutch tweets". *Computational Linguistics in the Netherlands Journal*, vol. 4, pp. 171-190, 2019.

[18] H. D. Majeed and G. S. Nariman. "Offline handwritten English alphabet recognition (OHEAR)". *UHD Journal of Science and Technology*, vol. 6, no. 2, pp. 29-38, 2022.

[19] K. Todorov and G. Colavizza. "An Assessment of the Impact of OCR Noise on Language Models". In: Proceedings of the 14th International Conference on Agents and Artificial Intelligence, 2022.

[20] M. Del Buono, L. Boatto, V. Consorti, V. Eramo, A. Esposito, F. Melcarne and M. Tucci. "Recognition of Handprinted Characters in Italian Cadastral Maps". In: Character Recognition Technologies. SPIE Proceedings, 1993. vol. 1906, pp. 89-99.

[21] R. Barman, M. Ehrmann, S. Clematide, S. A. Oliveira and F. Kaplan, "Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining and Digital Humanities*, 2021.

[22] F. Lopes, C. Teixeira and H. G. Oliveira. "Comparing different methods for named entity recognition in Portuguese neurology text". *Journal of Medical Systems*, vol. 44, no. 4, p. 77, 2020.

[23] N. Alrasheed, P. Rao and V. Grieco. "Character Recognition of seventeenth-century Spanish American notary records using deep learning". *Digital Humanities Quarterly*, vol. 15, no. 4, 2021.

[24] T. Q. Vinh, L. H. Duy and N. T. Nhan. "Vietnamese handwritten character recognition using convolutional neural network". *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 276-283, 2020.

[25] A. Chaudhuri, K. Mandaviya, P. Badelia and S. K. Ghosh. "Optical character recognition systems for German language." In: Optical Character Recognition Systems for Different Languages with Soft Computing. Cham, Springer International Publishing, 2017, pp. 137-164.

[26] D. Gunawan, D. Arisandi, F. M. Ginting, R. F. Rahmat and A. Amalia. "Russian character recognition using self-organizing map". *Journal of Physics*: *Conference Series*, vol. 801, p. 012040, 2017.

[27] G. Georgiev, P. Nakov, K. Ganchev, P. Osenova and K. I. Simov. "Feature-rich Named Entity Recognition for Bulgarian using Conditional Random Fields". In: Proceedings of the International Conference RANLP-2009. arXiv [cs.CL], 2021.

[28] A. Radchenko, R. Zarovsky and V. Kazymyr, "Method of Segmentation and Recognition of Ukrainian License Plates". In: 2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF), 2017.

[29] M. Gjoreski, G. Zajkovski, A. Bogatinov, G. Madjarov, D. Gjorgjevikj and H. Gjoreski. "Optical Character Recognition Applied on Receipts Printed in Macedonian Language". In: International Conference on Informatics and Information Technologies (CIIT), 2014.

[30] T. Ghukasyan, G. Davtyan, K. Avetisyan and I. Andrianov. "PioNER: Datasets and Baselines for Armenian Named Entity Recognition". In: 2018 Ivannikov Ispras Open Conference (ISPRAS), 2018.

[31] N. Alrobah and S. Albahli. "Arabic handwritten recognition using deep learning: A survey". *Arabian Journal for Science and Engineering*, 2022.

[32] O. Keren, T. Avinari, R. Tsarfaty and O. Levy, "Breaking Character: Are Subwords Good Enough for MRLs after all?" arXiv [cs.CL], 2022.

[33] Y. A. Nanehkaran, D. Zhang, S. Salimi, J. Chen, Y. Tian and N. Al-Nabhan. "Analysis and comparison of machine learning classifiers and deep neural networks techniques for recognition of Farsi handwritten digits". *Journal of Supercomputing*, vol. 77, no. 4, pp. 3193-3222, 2021.

[34] D. Rashid and N. Kumar Gondhi. "Scrutinization of Urdu handwritten text recognition with machine learning approach". In: Communications in Computer and Information Science. Cham, Springer International Publishing, 2022, pp. 383-394.

[35] Y. Wang, H. Mamat, X. Xu, A. Aysa and K. Ubul. Scene Uyghur text detection based on fine-grained feature representation". *Sensors* (*Basel*), vol. 22, no. 12, p. 4372, 2022.

[36] S. Sharma and S. Gupta. "Recognition of various scripts using machine learning and deep learning techniques-A review". In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021.

[37] P. D. Doshi and P. A. Vanjara. "A Comprehensive survey on Handwritten Gujarati Character and its Modifier Recognition Methods". In: Information and Communication Technology for Competitive Strategies (ICTCS 2020). Springer Singapore, Singapore, 2022, pp. 841-850.

[38] M. R. Haque, M. G. Azam, S. M. Milon, M. S. Hossain, M. A. A. Molla and M. S. Uddin. "Quantitative Analysis of deep CNNs for Multilingual Handwritten Digit Recognition". In: Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2021, pp. 15-25.

[39] H. Singh, R. K. Sharma and V. P. Singh. "Online handwriting recognition systems for Indic and non-Indic scripts: A review". *Artificial Intelligence Review*, vol. 54, no. 2, pp. 1525-1579, 2021.

[40] L. Saysourinhong, B. Zhu and M. Nakagawa. "Online handwritten Lao character recognition by MRF". *IEICE Transactions on Information and Systems*, vol. E95.D, no. 6, pp. 1603-1609, 2012.

[41] C. S. Lwin and W. Xiangqian. "Myanmar Handwritten Character Recognition from Similar Character Groups using K-means and Convolutional Neural Network". In: 2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE), 2020.

[42] M. A. Rasyidi, T. Bariyah, Y. I. Riskajaya and A. D. Septyani. "Classification of handwritten Javanese script using random forest

algorithm”. *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1308-1315, 2021.

[43] I. W. A. Darma and N. K. Ariasih. “Handwritten Balinesse Character Recognition using K-Nearest Neighbor”. INA-Rxiv, 2018.

[44] J. Park, E. Lee, Y. Kim, I. Kang, H. I. Koo and N. I. Cho. “Multilingual optical character recognition system using the reinforcement learning of character segmenter”. *IEEE Access*, vol. 8, pp. 174437-174448, 2020.

[45] R. Plamondon and S. N. Srihari. “Online and off-line handwriting recognition: A comprehensive survey*”. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, 2000.

[46] N. S. Guptha, V. Balamurugan, G. Megharaj, K. N. A. Sattar and J. D. Rose, “Cross lingual handwritten character recognition using long short term memory network with aid of elephant herding optimization algorithm”. *Pattern Recognition Letters*, vol. 159, pp. 16-22, 2022.

[47] G. S. Katkar and M. V Kapoor. “Performance analysis of structure similarity algorithm for the recognition of printed cursive English alphabets”. *International Journal of Scientific Research in Science and Technology*, vol.8, no.5, pp. 555-559, 2021.

[48] S. Tabassum, N. Abedin, M. M. Rahman, M. M. Rahman, M. T. Ahmed, R. I. Maruf and A. Ahmed. “An online cursive handwritten medical words recognition system for busy doctors in developing countries for ensuring efficient healthcare service delivery”. *Scientific Reports*, vol. 12, no. 1, p. 3601, 2022.

[49] D. H. Wang, C. L. Liu, J. L. Yu and X. D. Zhou. “CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters”. In: 2009 10th International Conference on Document Analysis and Recognition, 2009.

[50] T. Q. Wang, X. Jiang and C. L. Liu. “Query pixel guided stroke extraction with model-based matching for offline handwritten Chinese characters”. *Pattern Recognition*, vol. 123, p. 108416, 2022.

[51] A. Qaroush, B. Jaber, K. Mohammad, M. Washaha, E. Maali and N. Nayef. “An efficient, font independent word and character segmentation algorithm for printed Arabic text”. *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 1, pp. 1330-1344, 2022.

[52] K. M. M. Yaagoup and M. E. M. Musa. “Online Arabic handwriting characters recognition using deep learning”. *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 9, no. 10, pp. 83-92, 2020.

[53] P. B. Pati, S. Sabari Raju, N. Pati and A. G. Ramakrishnan. “Gabor Filters for Document Analysis in Indian Bilingual Documents.” In: International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of, 2004, pp. 123-126

[54] S. M. Obaidullah, C. Halder, N. Das sand K. Roy. “Numeral script identification from handwritten document images”. *Procedia Computer Science*, vol. 54, pp. 585-594, 2015.

[55] R. Bashir and S. Quadri. “Identification of Kashmiri Script in a Bilingual Document Image”. In: 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), 2013.

[56] S. Manjula and R. S. Hegadi. “Identification and Classification of Multilingual Document using Maximized Mutual Information”. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[57] K. Roy, O. M. Sk, C. Halder, K. Santosh and N. Das. “Automatic line-level script identification from handwritten document images-a region-wise classification framework for Indian subcontinent”. *Malaysian Journal of Computer Science*, vol. 31, no. 1, p. 10, 2016.

[58] G.S. Rao, M. Imanuddin and B. Harikumar. “Script Identification of Telugu, English and Hindi document image”. *International Journal of Advanced Engineering and Global Technology*, vol. 2, no. 2, pp. 443-452, 2014.

[59] E. O. Omayio, I. Sreedevi and J. Panda. “Word Segmentation by Component Tracing and Association (CTA) Technique”. *Journal of Engineering Research*, 2022.

[60] P. K. Singh, R. Sarkar and M. Nasipuri. “Offline script identification from multilingual Indic-script documents: A state-of-the-art”. *Computer Science Review*, vol. 15-16, pp. 1-28, 2015.

[61] Y. Baek, D. Nam, S. Park, J. Lee, S. Shin, J. Baek, C. Y. Lee and H. Lee. “CLEval: Character-level Evaluation for Text Detection and Recognition Tasks”. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.

[62] K. J. Taher and H. D. Majeed. “Recognition of handwritten English numerals based on combining structural and statistical features”. *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, vol. 21, no. 1, pp. 73-83, 2021.

[63] 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.

[64] Sakshi and V. Kukreja. “A retrospective study on handwritten mathematical symbols and expressions: Classification and recognition”. E*ngineering Applications of Artificial Intelligence*, vol. 103, p. 104292, 2021.

[65] N. Murugan, R. Sivakumar, G. Yukesh and J. Vishnupriyan. “Recognition of Character from Handwritten”. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 1417-1419.

[66] R. Parthiban, R. Ezhilarasi and D. Saravanan. “Optical Character Recognition for English Handwritten Text using Recurrent Neural Network”. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020.

[67] H. Q. Ung, C. T. Nguyen, K. M. Phan, V. T. M. Khuong and M. Nakagawa. “Clustering online handwritten mathematical expressions”. *Pattern Recognition Letters*, vol. 146, pp. 267-275, 2021.

[68] N. Gautam and S. S. Chai. “Zig-zag diagonal and ANN for English character recognition”. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.4, pp. 57-62, 2019.

[69] L. Deng. “The MNIST database of handwritten digit images for machine learning research [best of the web]”. *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141-142, 2012.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Kurdish Kurmanji Lemmatization and Spell-checker with Spell-correction

Hanar Hoshyar Mustafa, Rebwar M. Nabi

Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq

ABSTRACT

There are many studies about using lemmatization and spell-checker with spell-correction regarding English, Arabic, and Persian languages but only few studies found regarding low-resource languages such as Kurdish language and more specifically for Kurmanji dialect, which increased the need of creating such systems. Lemmatization is the process of determining a base or dictionary form (lemma) for a specific surface pattern, whereas spell-checkers and spell-correctors determine whether a word is correctly spelled also correct a range of spelling errors, respectively. This research aims to present a lemmatization and a word-level error correction system for Kurdish Kurmanji Dialect, which are the first tools for this dialect based on our knowledge. The proposed approach for lemmatization is built on morphological rules, and a hybrid approach that relies on the n-gram language model and the Jaccard Coefficient Similarity algorithm was applied to the spell-checker and spell-correction. The process results for lemmatization, as detailed in this article, rates of 97.7% and 99.3% accuracy for noun and verb lemmatization, correspondingly. Furthermore, for spell-checker and spell-correction, accordingly, accuracy rates of 100% and 90.77% are attained.

Index Terms: Kurdish Language, Kurmanji Dialect, Kurdish Lemmatizer, Kurdish Spell-checker and Spell-correction, Kurdish Dataset

1. INTRODUCTION

The topic of text processing has drawn the interest of numerous scholars as a result of the rising prevalence of digital texts in modern life. The amount of research in the domain of Kurdish text processing seems to be rather minor, despite significant efforts with some of the most popular languages, such as English, Persian, and Arabic.

Commonly, the language experts divided the used languages of the world over families which are by ascending: Indo-European, Sino-Tibetan, Niger-Congo, Austronesian, and some other families. The Indo-European family is the biggest family which speaks by the majority of Europe, the lands where the Europeans migrated, as well as a large portion of South-west and South Asia. This family divided into sub-families [1]. Kurdish language dialects are part of the north-western branch of the Indo-Iranic language family. The Kurdish language is an independent language that has its own linguistic continuum, historical origins, grammar rules, and extensive live linguistic skills. The “Median” or “Proto-Kurdish” language is where the Kurdish language originated. Approximately 30 million people in high land of Middle East, Kurdistan, talk numerous dialects of Kurdish [1].

Kurdish is referred to be a dialectical continuity, which means that it has a variety of dialects, it actually has four primary dialects (groups) and sub dialects, including (Kurmanjí or Kurmanji Zhwrw and Badínaní) in the north of Kurdistan and Sorani or Kurmanji Khwarw in the center

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp43-52 | E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author’s e-mail: Hanar Hoshyar Mustafa, Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq. E-mail: hanar.hoshyar.m@spu.edu.iq

Received: 23-10-2022 Accepted: 05-01-2023 Published: 22-02-2023

of Kurdistan (Sulaimani and Mukrayani). Kurmanji and Sorani are indeed the two main dialects [2]. Additionally, the other two important divisions of Kurdish language are Goraní (Hawrami, Zazayee and Shabak) and Luri (Mamasani, Kurmanshani and Kalhuri). Furthermore, these are categorized into dozens of dialects and sub-dialects [3]. This paper focuses on the Northern Kurdish dialect which is (Kurmanji or Kurmanji Zhwrw) dialect which has the biggest number of speakers in comparison to other Kurdish languages dialects [4]. Several studies have been done related to common languages such English [5], [6], Arabic [7]-[9], and Persian [10]-[12]. Moreover, there are few studies which are consummated regarding Kurdish language [13], [14], despite it, a huge gap can be seen in the case of Kurdish Kurmanji dialect; therefore, this study has been aimed to serve this gap due to Kurmanji dialect in the case of creating lemmatization and spell-checker with spell-correction system. Hence, in the future, this study can be used in several applications that include data translation, sentence retrieval, document retrieval, and also can be extend and upgrade to more powerful similar systems.

This study presented a toolkit, which consists of a lemmatization system and a spell-checker with spell-correction for Kurdish Kurmanji. The aim of the lemmatization is to find a root or dictionary form (calls a lemma) for a specific surface form. It is crucial to be able to normalize words into their most basic forms, particularly for languages with rich morphology such as Kurdish language, to better assist processes such as search engines and linguistic case studies.

Spell-checking algorithms are one of the lemmatizer's most commonly used applications. With using a spell checker, the system suggests a rating of suggested corrections for each possibly incorrect word.

This study presented a combination algorithm which are n-gram language model together with Jaccard Similarity Coefficient for the spell-checker and spell-correction system. Furthermore, a rule-based method on the Kurdish Kurmanji morphological rules is used in creating the lemmatization system.

Based on the literature and to the best of our knowledge, no study has been conducted regarding the spell-checking and lemmatization systems in Kurdish Kurmanji Dialect. Therefore, our study can be the base for further studies for Kurdish Kurmanji dialect.

2. RELATED WORK

There has been a huge amount of research that has been conducted regarding the word lemmatization, spell-checker, and spell-correction in several common languages, such as English, Persian, and Arabic. However, when it comes to Kurdish language, a large absence can be observed, especially in lemmatization and spell-checking with spell-correction system in Kurdish Kurmanji dialect.

In the case of lemmatizer in English language Lemma Chase which is a lemmatizer is created [5] address the problems of the most widely used lemmatizers currently available, this research presents a lemmatization model. This model accounts for the nominalized/derived terms for which no lemmatizer currently in use is able to produce the proper lemmas. Identifying the morphological structure of any input English word, and in particular understanding the structure of the derivational word, is the main issue in developing a lemmatizer. Finding the derivational suffix from morphing words and then extracting the dictionary base word from that derived word is another crucially difficult problem for a lemmatizer. Some derivative terms are not handled by well-known and well-liked lemmatizers to retrieve their basis words. Lemma Chase, the mentioned lemmatizer, accurately retrieves the base word while taking into account the word's Part of Speech, several classes of suffix rules, and effectively executing the recoding rules utilizing the WordNet Dictionary. All of the derivational and nominalized word forms that are present in any standard English dictionary are successfully used by Lemma Chase to construct the base word form.

In addition, there have been numerous studies on spell checkers in Arabic. For instance, Build Fast and Accurate Lemmatization for Arabic [7] which is a study that covers the need for a quick and precise lammatization to improve Arabic Information Retrieval (IR) outcomes and the difficulty of developing a lemmatizer for Arabic, since it has a rich and complex derivational morphology. Introduces a new data set that can be used to verify lemmatization accuracy as well as a powerful lemmatization algorithm that works more accurately and quickly than current Arabic lemmatization techniques.

Numerous studies have been published on the use of spell checkers and spell correction in Persian as well. For example, Automated Misspelling Detection and Correction in Persian Clinical Text [10] is an article that explains the creation of an automatic method for identifying and fixing misspellings in Persian free texts related to radiology and ultrasound.

Three distinct forms of free texts associated to abdominal and pelvic ultrasound, head-and-neck ultrasound, and breast ultrasound reports are utilized using n-gram language model to accomplish their aim. For free texts in radiology and ultrasound, the system obtained detection performance of up to 90.29% with correction accuracy of 88.56%. The findings suggested that clinical reports can benefit from high-quality spelling correction. Significant cost reductions were also made by the system throughout the documentation and final approval of the reports in the imaging department.

Kurdish stemmer pre-processing for improving information retrieval conducted by researcher in [13]. This article introduces the Kurdish stemming-step method. It is a method that links search phrases and indexing terms in Kurdish texts that are connected by morphology. In actuality, the occurrence of words demonstrates a supportive role for the classification process. Even though it was planned to produce more or fewer errors to demonstrate the complexity and difficulty of words in the Kurdish Sorani dialect, the handling of similarity changes was implemented, which helped to boost matching among words and decrease the storage requirements. However, the stemmer used in this work was capable of resolving most of these issues. There are many stop words with added affixes in Kurdish Sorani writings. Therefore, by combining these commonly occurring stop words, it can be stemmed. In addition, it was determined that employing partial words during the pre-processing stage was preferable.

Likewise building a Lemmatizer and a Spell-checker for Kurdish Sorani presented by [14]. This study also presented a lemmatization and word-level error correction system for Kurdish Sorani. It suggested a hybrid strategy focused on n-gram language modeling and morphological principles. Systems for lemmatization and error detection are referred to as Peyv and Renus, respectively. The Peyv lemmatizer is created based on the morphological rules, and for Renus, it corrects words both with using a lexicon and without using a lexicon. It indicates that these two basic text processing methods can lead the way for more study on additional natural language processing applications for Kurdish Sorani.

Last but not least, intensive literature search has been conducted but no studies have been found considering the Kurdish Kurmanji Dialect. Therefore, this article's primary goal is to propose a lemmatization and word-level spell checker with correction method for a Kurdish language dialect known as Kurmanji. The benchmark of this paper is [14] which is useful for the research study, despite the different algorithms used in spell-correction tool, the lemmatization tools are nearly similar in using the methods and approaches, both studies suggest a hybrid strategy based on n-gram language model and morphological principles. This study employs the Python programming language to process data as well as to create a word processing system that performs lemmatization and spell checking with spell correction at the word level.

3. METHODS AND DATA

This section describes dataset collection, data preparation, and algorithms as well as approaches which have been used in lemmatization and spell checker.

3.1. Dataset Collection

A model dataset was produced in order to carry out this study. The dataset was created by reading books and articles written in the Kurdish Kurmanji dialect, which were then manually recorded and added to the dataset. Kurdish Kurmanji dialect words include verbs, nouns, conjunctions, stop words, pronouns, imperative words, superlative words, and question words. There are around 1200 words in the dataset. Fig. 1 depicts the dataset's data amounts in a pie chart. This split results from the differing morphological rules for nouns and verbs, which affect how nouns and verbs are lemmatized. The third dataset has a large number of words that do not accept any affixes. Furthermore, it contains a few special terms with only one or two letters. Some of the conjunction words, for instance, are written with only one or two letters.

3.2. Data Preparation

The most important features that indicated that the dataset was ready for analysis were its unity and quality. Furthermore,

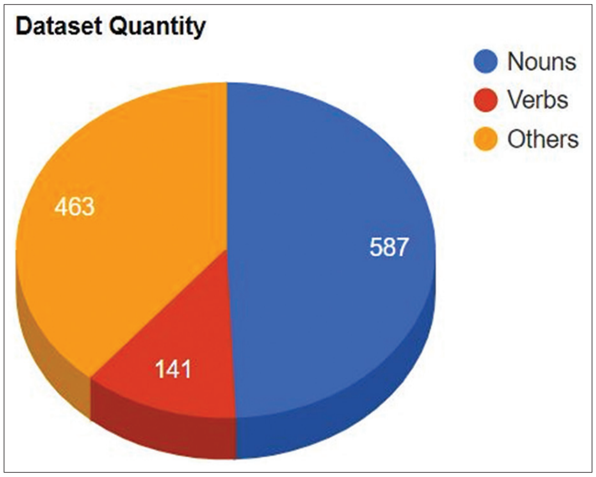


Fig. 1. Dataset quantity pie chart.

because the dataset is the primary first-hand collected dataset, it can ensure that the dataset is clean and has no duplicates. The dataset is then divided into three subsets. The first subset includes nouns. The second subset includes verbs, while the third subset contains pronouns, stop words, conjunctions, imperative words, superlative words, and question words. All of the subsets were stored in separate Excel files, each with two columns: the ID column and the data (word) column. Except for the third subset, which contains the verbs, it has four columns: ID, Chawg, Qad, and Rag. The ID column contains a unique ID for each row; the Chawg column contains the verb's base; the Qad column contains the verb's past root; and the Rag column contains the verb's present root. Table 1 presents the structure of the third (verb) Excel sheet.

3.3. Implementation

This section describes the approaches and methods used according to noun lemmatization, verb lemmatization and spell-checker.

3.3.1. Lemmatization

Lemmatizations for nouns and verbs are developed separately, after obtaining the fundamental morphological rules in Kurdish Kurmanji. Each of noun and verb lemmatization use different approaches based on the morphological rules. For the lemmatizations a pruning method is used to find out the root of the input word. In the background of the system, each process is contained in a module inside the system, as a result to eliminate complexity and increase simplicity, also to made the system more readable and understandable.

The following subsections clarify each of noun and verb lemmatizations in detail.

3.3.1.1. Noun lemmatization

According to the noun lemmatization, the noun lemmatization was created after clarifying and writing down all the rules in accordance with nouns in Kurdish Kurmanji dialect. A pruning method is used in this study. The input word to the system went through multiple stages and processes until the system found the proper root for the input noun, which is called a lemma in lemmatization process.

TABLE 1: Structure of verb-dataset

Verb dataset	Column	Include
	ID Chawg Qad Rag	Data ID Base of verb Past root of verb Present root of verb

During the process of noun lemmatization, predefined affixes and nouns in the dataset are used to find a proper lemma for an input noun. The only condition is to enter the word with the correct spelling.

When a noun was entered, a search algorithm was used to look for it in the dataset. If the entered noun was a root without any affixes, the system determined that the input was correct and that no further processing was necessary. The output word in the outcome would be the base of the entered word. Fig. 2 shows the flowchart diagram of this process.

In other cases when the entered noun is with or attached to some affixes, in this study in the noun lemmatization module, three sets of affixes were defined. First set included prefixes that write before the noun without attaching to the noun directly, in Kurdish Kurmanji, there are some prefixes that write with a space separated with the noun. Second set included the prefixes which are write and attached directly to the beginning of the noun without any space. Moreover, the last set included suffixes which are directly attached to the end of the noun.

The entered noun went through multiple processes to find the root out. The system first removes any prefixes which are

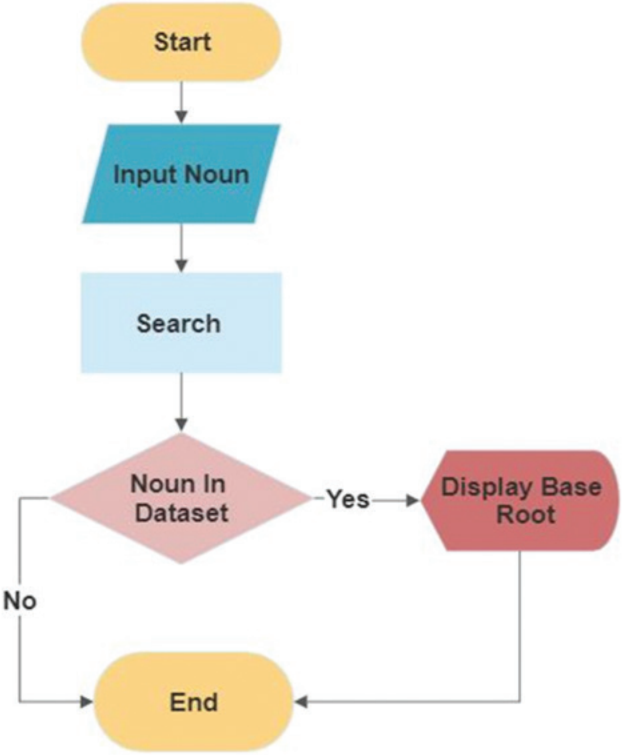


Fig. 2. Noun lemmatization first process flowchart.

attached or not attached prefixes to the word, then search in the dataset, if there was no matching for the entered noun, the system decided that it might attached to some suffixes too, then the word went through another process which removed the possible suffixes attached to the noun, after that a search process look to find out if there was any matching word in the dataset, if any matching word found in the dataset, it would be the return root as the result. This process showed in a flowchart diagram in Fig. 3.

Although there were no words that matched, the system made an effort and forwarded the entered noun to a procedure designed to remove prefixes and suffixes one at a time. In another sense, it took away the first prefix attached and looked for a matching root; if no matching root was discovered, it took away the first suffix attached and looked once more. It continued the process until the root was discovered if a matching root had not yet been discovered and there were further prefixes and suffixes linked to the word. At the end, when there were no more affixes, the entered noun was well spelled and the noun root existed in the dataset, the system gave the correct output lemma (root) for the entered noun. However, the system would replay with the message "Input word is not in the dataset" if there was no match between the entered noun and nouns in the dataset. Fig. 4. shows the process' flowchart diagram.

Following these steps, the user sees the procedures' output, as depicted in Figs. 5 and 6. In Fig. 5, the true word (کچ) (kiç) which means (girl) with two Kurdish Kurmanji prefixes (ەک) (ek) and (ا) (a) in the form of (کچەکا) (kiçeka) means (The girl who) entered. The system replayed with ("found", "کچ"); "found" denotes that the entered word is correct and already exists in the dataset, and "کچ" is the base root of word (کچەکا). However, in Fig. 6, the user inputted the incorrect term (کجان) (kican) in the meaning of (کچان) (kiçan) (girls) but with incorrect ending of (ج) (c) rather than (چ) (ç), followed by a correct prefix (ان) (an). Due to the incorrect spelling of the word, which confounded the system and prevented it from locating the specific base root of the word, the system replayed with the message "Input word is not in the dataset."

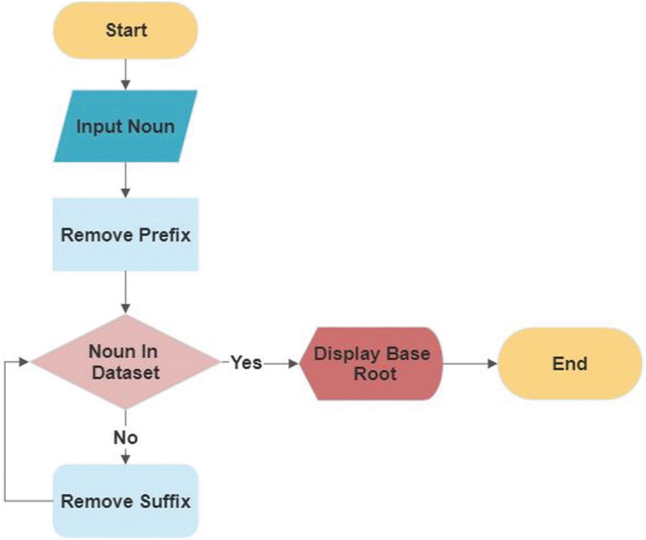


Fig. 3. Noun Lemmatization second process flowchart, Phase 1.

3.3.1.2. Verb lemmatization

Verb lemmatization also implemented in a pruning method as the noun lemmatization. After Kurdish Kurmanji dialect verb morphological rules are defined, the verb lemmatization is applied. The input verb went across several procedures until the tool selected and found the proper root.

Due to the Kurdish verb's morphology, the addition of prefixes and suffixes to the verb roots, and their ability to alter meaning, finding the root of the verb during the lemmatization process is more difficult and different than finding the root of a noun. Therefore, simply omitting the suffix is worthless.

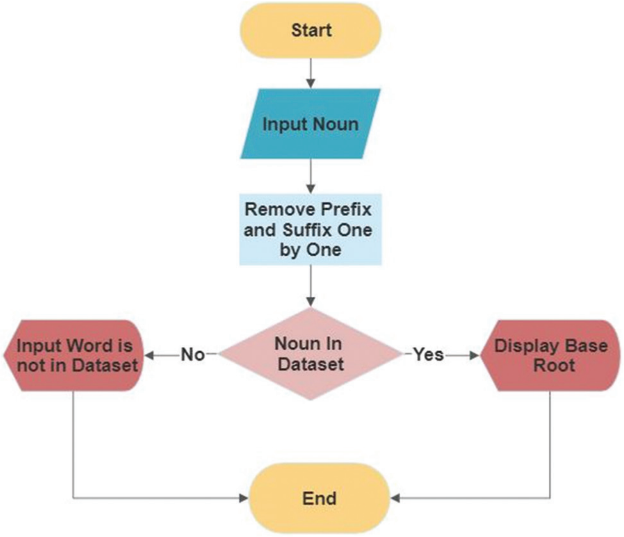
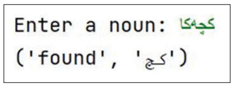
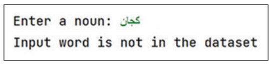

Start
Input Noun
Remove Prefix and Suffix One by One
Noun In Dataset
No
Yes
Input Word is not in Dataset
Display Base Root
End

Fig. 4. Noun Lemmatization second process flowchart, Phase 2.



```
Enter a noun: کچەکا
('found', 'کج')
```

Fig. 5. Noun Lemmatization of a legitimate noun.



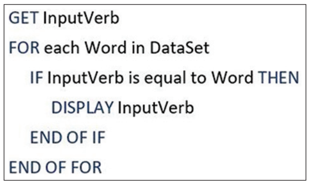
```
Enter a noun: کجان
Input word is not in the dataset
```

Fig. 6. Noun Lemmatization of an incorrect spelled noun.

In Kurdish language morphology, each verb has three states includes its critical state, which is called (Chawg) in Kurdish morphology; in this state, every verb ends with an (N) (ن) letter at the end of the word; the (N) (ن) is called (the N of Chawg) that determines the critical state of the verb. Another state is when the verb turns into its past state, which is called the "past root," and this is done by removing the (N of Chawg) at the end of the verb. Whenever the verb is in the past state, it can be used in the past tense. The final state is present, and it has several rules to modify a verb critical state and turn it into its present root. When the verb is changed to its present root, it can be used in the present tense [2].

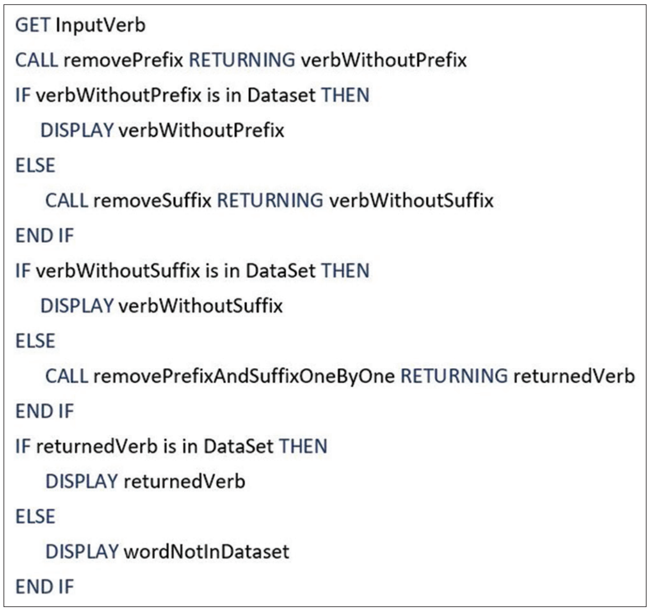
When the input was processed by the system, any affix containing the verb had to be removed. As a result, three sets of affixes are defined, which include suffixes, prefixes that do not attach to the verb, and prefixes that attach to the verb directly. After removing affixes, the remaining verb had to be compared with the verb dataset in the system. As it is clarified in the verb dataset excel file, there were four columns included (ID, Chawg, Qad, and Rag), in which Chawg referred to the critical state of the verb, Qad referred to the past state, and Rag referred to the present state. After the state of the verb was recognized and found, the system returned the critical state, which is the Chawg of the verb as the base root of the entered verb. Fig. 7 shows the process of finding the root of a verb if the entered verb is already a root; no matter in which tense it appears, the system returns the base root of it. Moreover, Fig. 8 depicts the processes for locating a verb root if the entered verb is attached to some affixes; the processes are identical to those for locating a root of a noun attached to affixes in noun lemmatization.

After completing these stages, the user sees the output of the procedures, as shown in Figs. 9-11. In Fig. 9, the true word (دخۆم) (dixom) denotes the present tense of the verb (خارن) (xarin) (eat), while the prefix (د) (d) indicates the present term of the verb and the suffix (م) (m) is the pronoun that denotes (I). The system repeated ("found", "خارن") in the output, where "found" implies that the word is correctly spelled and that its present root, which is (خۆ) (xo), is available in the dataset, and "خارن" is the base root for the entered word. In addition, in Fig. 10, the past tense of the same word (خارن) (eat) is entered



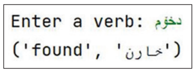
```
GET InputVerb
FOR each Word in DataSet
    IF InputVerb is equal to Word THEN
        DISPLAY InputVerb
    END OF IF
END OF FOR
```

Fig. 7. Verb Lemmatization first process pseudo code.



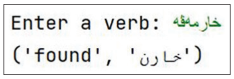
```
GET InputVerb
CALL removePrefix RETURNING verbWithoutPrefix
IF verbWithoutPrefix is in Dataset THEN
    DISPLAY verbWithoutPrefix
ELSE
    CALL removeSuffix RETURNING verbWithoutSuffix
END IF
IF verbWithoutSuffix is in DataSet THEN
    DISPLAY verbWithoutSuffix
ELSE
    CALL removePrefixAndSuffixOneByOne RETURNING returnedVerb
END IF
IF returnedVerb is in DataSet THEN
    DISPLAY returnedVerb
ELSE
    DISPLAY wordNotInDataset
END IF
```

Fig. 8. Verb Lemmatization second process pseudo code.



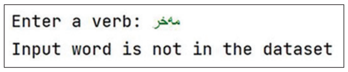
```
Enter a verb: دخۆم
('found', 'خارن')
```

Fig. 9. Verb Lemmatization of correct present tense of verb (خارن) (xarin) (eat).



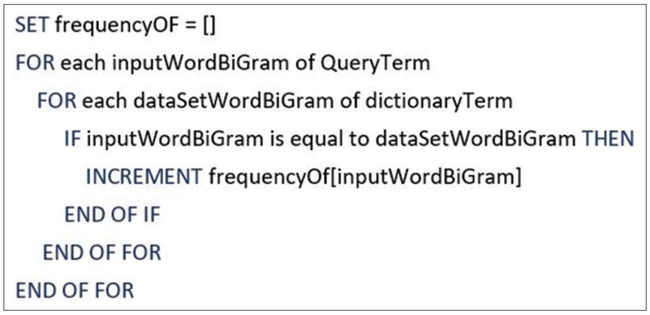
```
Enter a verb: خارمەڤە
('found', 'خارن')
```

Fig. 10. Verb Lemmatization of true past tense of verb (خارن) (xarin) (eat).



```
Enter a verb: مەخر
Input word is not in the dataset
```

Fig. 11. Verb Lemmatization of a wrong spelled negative imperative of verb (خارن) (xarin) (eat).



```
SET frequencyOF = []
FOR each inputWordBiGram of QueryTerm
    FOR each dataSetWordBiGram of dictionaryTerm
        IF inputWordBiGram is equal to dataSetWordBiGram THEN
            INCREMENT frequencyOf[inputWordBiGram]
        END OF IF
    END OF FOR
END OF FOR
```

Fig. 12. Query term bi-gram frequency calculation pseudo code.

as (خارمەڤە) (xarmeve), which means (I ate). This time, there are two suffixes: (م) (m), which is the pronoun associated to (I), and (ەڤە) (eve), which indicates that the event occurred and ended completely in the past. Once more, the system verified that the word root was correctly spelled that it was included in the dataset; it also displayed the base root of the term. Furthermore, in Fig. 11, entered the wrong negative imperative phrase (مەخر) (mexir) instead of (مەخو) (mexu) or (مەخۆ) (mexo) which means (don't eat), but with the improper ending of (ر) (r), rather than (ۆ) (o) or (و) (u). The system displayed the message "Input word is not in the dataset" according to the word's incorrect spelling, which confused the system and prohibited it from finding the precise base root of the word.

3.3.2. Spell checker and spell correction

The spell checker and spell correction mechanisms collaborated in two stages in this study: First, the spell checker indicated whether the word was correct or incorrect, and second, the spell correction process corrected the word by suggesting some correct words by providing the most likely correct word forms.

After the word entered the system, it was detected if it was true or not by the spell checker's check for word frequency in the dataset (including the whole of the three files). The step of finding that the word is true or detecting the word as wrong was done based on using n-grams. The input word, which is called the query term in this paper, is fragmented into bi-grams (two grammatical units). A bi-gram is an n-gram for $n = 2$. In this study, a 2-g (or bi-gram) is a two-letter sequence of letters. The bi-grams sequences "ha," "ap," "pp," and "py," for instance, are two-letter grammatical sequences extracted from the word (happy). After the bi-gram of the query term is produced, the system calculates the gram frequencies with the bi-grams of the words in the dataset separately, which is called a dictionary term in this paper. Fig. 12 shows the process of calculating the frequency of bi-grams in the query term in comparison to the dictionary terms. After calculation, the system looked up the frequencies of the bi-gram of the query term; if one of the frequencies was equal to zero, then it detected the word as a wrong one as one of its bi-grams had no repetition in comparison with the dictionary terms, and if none of the frequencies were zero, then the word was detected as true. Hence, in the event that a query term equals one of the index terms in the dataset, this word will be selected as true, and if the word is detected as true, then the system presents "The word is true spelled" as a result.

After detecting the query term as wrong, its bi-grams are handled, and the system goes to the spell correction procedure. The wrong word is then corrected based on the Jaccard similarity coefficient method, which is popularly used to compare how close the query terms in the dataset are to one another. Here, the procedure of similarity measurement can be used to examine the most comparable terms that are structurally recorded in the dataset if a query does not match any index in the dataset. Using the Jaccard similarity coefficient [15], Equation (1) shows the rule of the Jaccard similarity coefficient.

$$\text{Jaccard sim } (A,B) = P(A\cap B)/(BUA) \tag{1}$$

Measuring the Jaccard similarity coefficient between two datasets is done by dividing the number of features that are common to all by the number of properties [15].

The mechanism worked on the query term, and dictionary terms included all three files of the dataset. The spell correction took the query term, looked for the matching dictionary term in file one, if it did not exist, then sent it to the lemmatization files, respectively, because it may be the root of a noun or a verb; also, it might be a noun or a verb with affixes, and the affixes should be removed as a result to check if it was spelled correctly or not. After checked process did go well in detail, if the word found, the system marked it as a true word. Otherwise, spell checker predicted words based on the dataset's three files, then it chose the best matching words based on the highest matching degree, which is calculated using the Jaccard Coefficient algorithm, and best matches were chosen if their matching degree were greater than the spell checker's threshold, and finally the five highest matching degree words were chosen. The threshold of this study is equal to 0.15. It has been chosen based on the accuracy of the guess for the correct word or the highest matching words in the dataset for the wrong query term. In Kurdish Kurmanji, there are words with three letters; if they are written incorrectly by missing a letter, they only have two letters. Hence, the threshold should be as small as possible to get a great and accurate result.

4. RESULTS AND DISCUSSION

This section presents the results of the algorithms in both lemmatization and spell checker tools. Also discuss the benchmarking with the benchmark study of the research.

4.1. Noun Lemmatization

To improve the efficiency and accuracy of the noun lemmatization tool, two random words were chosen with their derivatives which were nine derivatives of (چیا) (mountain) word and 12 derivatives of (کوڕ) (boy) word. The results of lemmatization process of both words were successfully giving correct root in both nine derivatives of first word and 12 derivatives in second word. To ensure the accuracy of noun lemmatization, another 66 random words with possible derivatives were chose and entered into the system; therefore, the noun lemmatization gave correct result in 63 cases out of 66, which means that the noun lemmatization algorithm had an accuracy of approximately 95.45% in lemmatizing words. Overall, the accuracy of the noun lemmatization process was approximately about 97.7%. Table 2 presents accuracy in noun lemmatization tool.

4.2. Verb Lemmatization

To evaluate the efficiency and accuracy of the verb lemmatization tool, two sets of random verb forms were tested with the tool. The test sets included different verb forms such as present and past tense, imperative and negative imperative, passive, and negative. Regarding the verb's existence in the dataset dictionary, the verb lemmatization tool found the correct root of the input verb. Each verb in the test set was entered with all possible derivations made with specific prefixes and suffixes. The first set included 171 different forms of different verbs. The lemmatization tool lemmatized 169 of them correctly; the wrongly lemmatized ones were due to the ordering of the dataset; in the case of imperative and negative imperative of a verb, the lemmatized verb Rag was coming before the purposed verb Rag, so the system took the first verb Rag before it reached the purposed one. For example, the Kurdish verb "send" has two forms: (ناردن) (nardin) and (نارتن) (nartin) and both have the same Rag (نێر) ("nêr"). If a user entered the imperative tense of this verb, which is (بنێره) (binêre), and expected to see the base root of (ناردن) (nardin) in the result, the system replays with the base root of (نارتن) (nartin) because it was recorded before the other form (ناردن) (nardin) in the dataset excel file. Moreover, it is due to the system that, when it finds a result, it stops without going to the other verbs in the dataset.

Moreover, it is due to the system, when it finds a result, the system stops and the result appears without going to the other data in the dataset. As a result, the accuracy of lemmatizing the first set was 98.83 percent. In the order of the other set, there were 131 different forms of different verbs with different tenses. Due to this set, the lemmatization tool lemmatized all of them, which means it gave the correct root for each of the forms. It can be said that with the two test sets, the verb lemmatization tool overall gave approximately 99.4 percent accuracy. Table 3 shows the accuracy of the verb lemmatization tool.

TABLE 2: Accuracy in noun lemmatization tool

Sets	True lemmatization	False lemmatization	Total	Accuracy (%)
1st set	21	0	21	100
2nd set	63	3	66	95.45
Total	84	3	87	97.7

4.3. Spell Checker and Spell Correction

According to calculate and analyze the accuracy of the spell-checker and spell-correction tool, the process of analyzation is more complex, due to connecting the spell-checker and spell-correction tool with the lemmatization tools. As described in the above section, there was three datasets, so the spell-checker and spell-correction accuracy should be calculated according to all the datasets. The mechanism as said is to first check if the input word is correct or not, and the spell-checker tool is tested with three groups of data which are consisted in the three datasets as well. These three groups included 100 words from first dataset file, 100 nouns from second dataset file, 100 verbs from third dataset file, respectively. The result always returned true which meant the input word spelling is correct, while the data existed in the dataset. Hence, it reached to be said that the spell-checker tool returned in all cases successfully. Table 4 shows the accuracy of spell-checker tool.

For the spell-correction tool a set of random words included noun, verb and others is tested, the contained nouns and verbs included all forms with prefixes and suffixes also simple noun and verbs without prefixes and suffixes. The result shows that whenever a bi-gram of the original correct word came in the input word, it was a higher chance to get the most correct word and most similar word as a result. The more bi-

TABLE 3: Accuracy in verb lemmatization tool

Sets	True lemmatization	False lemmatization	Total	Accuracy (%)
1st set	169	2	171	98.83
2nd set	131	0	131	100
Total	300	2	302	99.3

TABLE 4: Accuracy in spell checker tool

Sets	True spell checking	False spell checking	Total	Accuracy (%)
1st set	100	0	100	100
2nd set	100	0	100	100
3rd set	100	0	100	100
Total	300	0	0	100

TABLE 5: Accuracy in spell correction tool

Sets	True correction	False correction	Total	Accuracy (%)
1st set	55	6	61	90.16
2nd set	71	9	80	88.75
3rd set	100	7	107	93.4
Total	226	22	254	90.77

grams of the original wanted word came in the input word, the higher similarity degree get and the more accurate results acquire in the outcome. In several occasions, the incorrect lemmatization occurred because of the incorrect input word, and this led to incorrect spell-correction which at the end resulted in a low accuracy degree of the outcome result.

To provide the efficiency of the spell-correction a set included 100 of wrong random words with different forms were tested manually. First set included 61 wrong spelled nouns, the spell-corrector with the help of the noun lemmatization resulted an accuracy of 90.16% of the correction process. Second set contained 80 wrong spelled verbs, in the result spell-corrector with the use of verb lemmatization gave an accuracy of correction process with 88.75% rate. Third set consisted the wrong spelled pronouns, stop words, conjunctions, imperative words, superlative words, and question words, in 107 words the spell-correction system corrected 100 of them successfully which give accurate result as 93.4% of accuracy rate. Table 5 displays the accuracy of spell-correction tool.

As shown in Table 5, the third set had the highest accuracy rate among the other two sets, and as previously stated, some false correction cases occurred due to false lemmatization, so it must be stated that if a dataset is created with all the forms of the words in all three datasets, then more accurate results can be obtained because the spell-corrector can directly look for the right form of the input misspelled word and find it with a high degree of certainty.

5. CONCLUSION AND FUTURE WORKS

Information retrieval and text classification can benefit greatly from effective lemmatizer. In addition, incorrect words are detected and corrected by spell-checkers and spell-correction. This paper introduced the Kurdish Kurmanji lemmatizer and word-level spell-checker with spell-correction methodologies. It is the first attempt that tools of this kind have been made for Kurdish Kurmanji. A hybrid technique has been utilized for the spell-checker and spell-correction that depends on the n-gram language model and the Jaccard Coefficient Similarity algorithm, also the proposed approach for lemmatization, is based on morphological principles. The outcome demonstrated that, while applying the suggested approach, the accuracy of lemmatization for each noun and verb lemmatization was assessed, respectively, at 97.7% and 99.3%. In addition, the spell-checker and spell-correction accuracy rates were 100% and 90.77%, respectively. The experimental findings show that several false correction cases were caused by incorrect lemmatization led by misspelled input words. Furthermore, according to experimental findings, more accurate results may be obtained if a dataset is established with all the word forms in the datasets since the spell-checker will directly search for the correct form of the input misspelled word and discover it with a high level of equality. In the future, this work can be expanded to apply to a bigger dataset of Kurdish Kurmanji and utilize these approaches for NLP applications like text mining for Kurdish Kurmanji.

As a contrast between this study and its benchmark. Actually, this study is done for the Kurdish Kurmanji dialect, while the benchmark was done for the Kurdish Sorani dialect, which has completely different morphological rules in so many phases to study and implement in the system. The datasets that were used were different, while this research's dataset is primary, first-hand, and organized in three subsets. In addition, there are some variances between them in terms of accuracy and the algorithms that have been used. This study achieved 97.7% and 99.3% accuracy for noun and verb lemmatization, respectively, while the benchmark achieved 95% and 89.4% accuracy of two test sets for noun lemmatization and an average of 86.7% accuracy for verb lemmatization. In addition, according to the spell-correction, this study used the Jaccard Coefficient Similarity algorithm and rated 90.77% accuracy, while the other study, as mentioned, used an edit distance algorithm and obtained 96.4% accuracy with a lexicon while, without a lexicon, the correction system had 87% of accuracy. At the end, it has to be said that the similarities can be seen in the theoretical parts and ideas, but for the practical part, a huge difference can be seen from using different programming languages; this study used the Python programming language, while the other used the Java programming language, up to and including recreating the system from the beginning to the end.

6. ACKNOWLEDGMENT

The authors would like to thank SPU for providing the opportunity, support, and funding for this study. Sulaimani, the Kurdish journalist syndicate, is also thanked.

REFERENCES

[1] Z. Kurdî, M.Û. Zarên Wî and H.S. Khalid. “Kurdish Language, its Family and Dialects”. 2020. Available from: https://www.dergipark.org.tr/en/pub/kurdiname/issue/50233/637080 [Last accessed on 2022 Aug 15].

[2] D.N. MacKenzie. “Kurdish Dialect Studies”. Oxford University Press, London, 1961. Available from: https://www.books.google.iq/books/about/Kurdish_dialect_studies_2_1962.html?id=eaf2zaeacaaj&redir_esc=y [Last accessed on 2022 May 31]

[3] “Kurdish Academy of Language Enables the Kurdish Language in New Horizon”. Available from: https://www.kurdishacademy.org/?q=node/41 [Last accessed on 2022 Jun 04].

[4] N.A. Khoshnaw, Z.U.Z. Sulaimaniyah. “Awer Station”, 2011. Available from: https://rezmanikurde.blogspot.com/2018/01/blog-post_26.html?m=1 [Last accessed on 2022 Jun 09].

[5] R. Gupta and A.G. Jivani. “LemmaChase: A Lemmatizer”. *International Journal on Emerging Technologies*, vol. 11, no. 2, pp. 817-824, 2020.

[6] D. Hládek, J. Staš, S. Ondáš, J. Juhár and L. Kovács. “Learning string distance with smoothing for OCR spelling correction”. *Multimedia Tools and Applications*, vol. 76, no. 22, pp. 24549-24567, 2017.

[7] H. Mubarak. “Build Fast and Accurate Lemmatization for Arabic”. vol. Proceedings of the European Language Resources Association (ELRA). Miyazaki, Japan, 2018. Available from: https://www.aclanthology.org/L18-118 [Last accessed on 2022 Jun 08].

[8] N. Zukarnain, B.S. Abbas, S. Wayan, A. Trisetyarso and C.H. Kang. “Spelling Checker Algorithm Methods for Many Languages”, in Proceedings of 2019 International Conference on Information Management and Technology, (ICIMTech), 2019, pp. 198-201.

[9] A.A. Freihat, M. Abbas, G. Bella and F. Giunchiglia. “Towards an optimal solution to lemmatization in Arabic”. *Procedia Computer Science*, vol. 142, pp. 132-140, 2018.

[10] A. Yazdani, M. Ghazisaeedi, N. Ahmadinejad, M. Giti, H. Amjadi and A. Nahvijou. “Automated misspelling detection and correction in Persian clinical text”. *Journal of Digital Imaging*, vol. 33, no. 3, pp. 555-562. 2019.

[11] S. Mohtaj, B. Roshanfekr, A. Zafarian and H. Asghari, “Parsivar: A Language Processing Toolkit for Persian,” in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018. Available from: https://www.aclanthology.org/L18-1179 [Last accessed on 2022 Aug 20].

[12] A. Rashidi and M.Z. Lighvan. HPS: A hierarchical Persian stemming method. *International Journal on Natural Language Computing*, vol. 3, no. 1, pp. 11-20, 2014.

[13] A.M. Mustafa and T.A. Rashid. Kurdish stemmer pre-processing steps for improving information retrieval. *Journal of Information Science*, vol. 44, no. 1, pp. 15-27, 2018.

[14] S. Salavati and S. Ahmadi. “Building a Lemmatizer and a spell-checker for Sorani Kurdish”. CoRR, vol. abs/1809.10763, 2018. Available from: https://www.arxiv.org/abs/1809.10763 [Last accessed on 2021 Aug 15].

[15] S. Niwattanakul, J. Singthongcha, E. Naenudorn, and S. Wanapu. “Using of Jaccard Coefficient for Keywords Similarity”, in Proceedings of the International Multi Conference of Engineers and Computer Scientists. vol. 1, 2013. Available from: https://www.data.mendeley.com/v1/datasets/s9wyvvbj9j/draft?preview=1 [Last accessed on 2022 Apr 08].

REVIEW ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Review on IoT Intrusion Detection Systems Using Supervised Machine Learning: Techniques, Datasets, and Algorithms

Azeez Rahman Abdulla, Noor Ghazi M. Jameel

Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq

ABSTRACT

Physical objects that may communicate with one another are referred to "things" throughout the Internet of Things (IoT) concept. It introduces a variety of services and activities that are both available, trustworthy and essential for human life. The IoT necessitates multifaceted security measures that prioritize communication protected by confidentiality, integrity and authentication services; data inside sensor nodes are encrypted and the network is secured against interruptions and attacks. As a result, the issue of communication security in an IoT network needs to be solved. Even though the IoT network is protected by encryption and authentication, cyber-attacks are still possible. Consequently, it's crucial to have an intrusion detection system (IDS) technology. In this paper, common and potential security threats to the IoT environment are explored. Then, based on evaluating and contrasting recent studies in the field of IoT intrusion detection, a review regarding the IoT IDSs is offered with regard to the methodologies, datasets and machine learning (ML) algorithms. In This study, the strengths and limitations of recent IoT intrusion detection techniques are determined, recent datasets collected from real or simulated IoT environment are explored, high-performing ML methods are discovered, and the gap in recent studies is identified.

Index Terms: Internet of thing, Intrusion detection, Intrusion detection system techniques, Intrusion detection system datasets, Supervised machine learning

1. INTRODUCTION

A smart network called the Internet of Things (IoT) employs established protocols to link things to the Internet [1]. In an IoT network, smart tiny sensors join objects wirelessly. IoT devices can interact with one another without human involvement [2]. It uses distinctive addressing techniques to communicate, add more items and collaborate with them to develop new applications and services. Examples of IoT applications include smart environments, smart homes, and smart cities [3]. Thereby of the development of IoT applications, several obstacles have developed. One of these obstacles is IoT security that cannot be disregarded. IoT networks are subject to a range of malicious attacks because IoT devices can be accessed from anywhere over an unprotected network such as the Internet. The following security requirements should be considered when securing IoT environment:

- Confidentiality: IoT systems must ensure that unauthorized parties are prohibited from disclosing information [4].
- Integrity: Ensures that the messages must not have been modified in any manner [4].
- Availability: When data or resources are needed, they must be available [4]. Attackers can saturate a resource's bandwidth to degrade its availability.

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp53-65	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Azeez Rahman Abdulla, Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq. Azeez.rahman.a@spu.edu.iq

Received: 18-10-2022 Accepted: 23-12-2022 Published: 01-03-2023

- Authenticity: The word "authenticity" relates to the ability to prove one's identity. The system should be able to recognize the identity of the entity with whom it is communicating [5].
- Non-repudiation: This guarantees that nothing can be rejected. In an IoT context, a node cannot reject a message or piece of data that has already been sent to another node or a user [6].
- Data freshness: Ensures that no outdated messages are retransmitted by an attacker [7].

In the last few years, advancement in artificial intelligent (AI) such as machine learning (ML) techniques has been used to improve IoT intrusion detection system (IDS). Numerous studies as [8,9], reviewed and compared different applied ML algorithms and techniques through various datasets to validate the development of IoT IDSs. However, it's still not clear a recent dataset collected from IoT environment, and which ML model was more effective for building an efficient IoT IDS. Therefore, the current requirement is to do an up-to-date review to identify these critical points.

In this study, a survey of the IoT IDSs is given. This paper aims to further the knowledge in regard to IoT cyber attacks' characteristics (motivation and capabilities). Then, strengths and limitations of different categories of IDSs techniques (hybrid, anomaly-based, signature-based, and specification-based) are compared. Moreover, the study presents a review on the recent researches in the area of IoT intrusion detection using ML algorithms for IoT network based on the datasets, algorithms and evaluation metrics to identify the recent IoT dataset and the outperformed ML algorithm in terms of accuracy used for IoT intrusion detection.

The paper is structured as follows: In section 2, common cyber-attacks in IoT the environment are clarified. In section 3 the strengths and limitations of IoT intrusion detection techniques are discussed. Section 4 discussed, analyzed and compared recent IoT intrusion detection researches' performance metrics, datasets and supervised ML algorithms. Finally, section 5 illustrates the conclusions of the paper.

2. IoT CYBER ATTACKS

Recently, IoT has developed quickly, making it the fastest-growing enormous impact of technology on social interactions and workplace environments, including education, healthcare and commerce. This technology is used for storing the private data of people and businesses, for financial data transactions, for product development and for marketing. Due to the widespread adoption of linked devices in the IoT, there is a huge global demand for strong security. Millions or perhaps billions of connected devices and services are now available [10-13]. Every day, there are more risks and assaults have gotten more frequent and sophisticated. In addition, sophisticated technologies are becoming more readily available to potential attackers [14,15]. To realize its full potential, IoT must be secured against threats and weaknesses [16]. By maintaining the confidentiality and integrity of information about the object and making that information easily accessible whenever it is needed, security is the act of avoiding physical injury, unauthorized access, theft, or loss to the item [17]. To ensure IoT security, it is crucial to maintain the greatest inherent value of both tangible items (devices) and intangible ones (services, information and data). System risks and vulnerabilities must be identified in order to provide a comprehensive set of security criteria to assess if the security solution is secure against malicious assaults or not [18]. Attacks are performed to damage a system or obstruct regular operations by utilizing various strategies and tools to exploit vulnerabilities. Attackers launch attacks to achieve goals, either for their personal satisfaction or to exact revenge [19]. Common IoT cyber-attack types are:

- Physical attacks: These assaults tamper with hardware elements. Most IoT devices often operate in outdoor areas which are extremely vulnerable to the physical assaults [20].
- Attacks known as reconnaissance include the illegal identification of systems, services, or vulnerabilities. The scanning of network ports is an example of a reconnaissance attack [21].
- Denial-of-service (DoS): This type of attack aims to prevent the targeted users from accessing a computer or network resource. The majority of IoT devices are susceptible to resource enervation attacks due to their limited capacity for memory and compute resources [22].
- Access attacks happen when unauthorized users get access to networks or devices that they are not allowed to use. Two types of access assaults exist: The first is physical access, in which a hacker gains access to a real object. The second is using IP-connected devices for remote access [22].
- Attacks on privacy: IoT privacy protection has grown more difficult as a result of the volume of information that is readily accessible via remote access techniques [14].
- Cyber-crimes: Users and data are used for hedonistic activities including fraud, brand theft, identity theft, and

theft of intellectual property using internet and smart products [14,15,23].

- Destructive attacks: Space is exploited to cause widespread disturbance and property and human life loss. Terrorism and retaliation are two examples of damaging assaults.
- Supervisory Control and Data Acquisition (SCADA) Attacks: SCADA systems are connected to industrial IoT networks; they are active devices in real-time industrial networks, which allow the remote monitoring and control of processes, even when the devices are located in remote areas. The most specific and common types of SCADA attacks are eavesdropping, man-in-the middle, masquerading, and malware [24].

3. IoT INTRUSION DETECTION SYSTEM

Despite the investment and potential it holds, there are still issues that prevent IoT from becoming a widely utilized technology. The security challenges with IoT are thought to be solvable via intrusion detection, which has been established for more than 30 years. Intrusion detection is often a system (referred to as IDS) which consists of tools or methods that analyze system activity to find assaults or unauthorized access. An IDS typically comprises of sensors, and a tool to evaluate the data from these sensors. Efficient and accurate intrusion detection solutions are necessary in the IoT environment to identify various security risks [25].

3.1. IoT Intrusion Detection Types

IDS types can be categorized in a variety of ways, particularly IDS for IoT as the majority of them are still being studied. According to Das *et al.*, [26] the research distinguishes three types of IDS:

- Host-based IDS (HIDS): To keep an eye on the system's harmful or malicious activity, HIDS is connected to the server. Specifically, HIDS examines changes in file-to-file communication, network traffic, system calls, running processes, and application logs. This sort of IDS's drawback is that it can only identify attacks on the systems it supports.
- Network-based IDS (NIDS): NIDS analyzes network traffic for attack activities and identifies harmful behavior on network lines.
- Distributed IDS (DIDS): DIDS will have a large number of linked and dispersed IDSs for attack detection, incident monitoring and anomaly detection. To monitor and respond to outside actions, DIDS needs a central server with strong computing and orchestration capabilities.

3.2. IoT Intrusion Detection Techniques

There are four basic types or methodologies for deploying IoT intrusion detection.

- Anomaly based IDS in IoT.
  It uses anomaly based IDS to find intrusions and monitor abusive behavior. It employs a threshold to determine if this behavior is typical or abnormal. These IDSs have the ability to monitor a typical IoT network's activity and set a threshold. To detect abnormalities, the network's activity is compared to a threshold and any deviation from this number is considered abnormal [27]. Table 1 compares and contrasts the strength and limitations of several anomaly-based IDSs methodologies based on resource and energy usage, detection accuracy and speed.
- Signature based IDS in IoT
  Signature based detections compare the network's current activity to pre-defined attack patterns. Each signature is connected to a particular assault since signatures are originally established and stored on the IoT device. Signature based approaches are commonly used and require a signature for each assault [27]. The strengths and limitations of different signature based IDSs techniques have been presented and compared in Table 2 based on resource consumption, energy, detection accuracy, and speed.
- Specification based IDS in IoT
  Specification-based approaches detect intrusions when network behavior deviates from specification definitions. Therefore, specification-based detection has the same purpose of anomaly-based detection. However, there is one important difference between these methods: In specification-based approaches, a human expert should manually define the rules of each specification [36]. The main aspects of specification-based IDSs have been outlined and then compared in Table 3 based on resource consumption, energy, detection accuracy, and speed.
- Hybrid IDS in IoT
  Signature based IDS has a large usable capacity and limited number of attack detections while anomaly based IDS has a high false positive rate and significant computation costs. A hybrid technique was suggested to solve the flaws of both systems [42]. The main characteristics of hybrid IDSs have been defined and then compared in Table 4 based on resource consumption, energy, detection accuracy, and speed.

TABLE 1: Comparison of different anomaly based IDS techniques

Reference No.	Technique	Strength	Limitations
[28]	Utilizing a fusion based technique to decrease the damage caused by strikes.	● Low communication overhead	● High energy consumption
[29]	Detecting Wormhole attacks using node position and neighbor information.	● Low resource consumption ● Real time ● Energy efficient	● Only One type of attack can be detected
[30]	Detecting sinkhole attacks by analyzing the behavior of devices	● Detection accuracy is high	● Detect limited number of attacks
[31]	A lightweight technique for identifying normal and deviant behavior	● Lightweight implementation ● Detection accuracy is high	● High computational overhead
[32]	A request-response method's correlation functions are used to look for unusual network server activity	● Consuming modest resources ● Lightweight detection system	● High computational overhead

IDS: Intrusion detection system

TABLE 2: Comparison of different signature based IDS techniques

Reference No.	Technique	Strength	Limitations
[33]	Detecting network attacks by signature code in IP based ubiquitous sensor networks	● High detection accuracy ● Low energy and resource consumption	● Can detect limited number of intrusions
[34]	The pattern-matching engine is used to detect malicious nodes using auxiliary shifting and early decision techniques	● Low memory and computational complexity ● Maximum speed up	● Not real-time ● Can detect limited number of intrusions
[35]	Detection of malware signature detection using reversible sketch structure based on cloud.	● Fast ● Low communication consumption ● High detection accuracy	● High memory requirement ● Has a limited ability to identify assaults

IDS: Intrusion detection system

TABLE 3: Comparison of different specification based IDS techniques

Reference No.	Technique	Strength	Limitations
[37]	Mitigation of black hole attacks Using an effective strategy in routing protocol for low-power and lossy (RPL) Networks	● Low delay ● High detection accuracy of the infected node	● Only black hole attacks can be detected
[38]	Detecting internal attacks by designing a secure routing protocol based on reputation mechanism	● Detection accuracy is acceptable ● Low delay	● Needs skilled administration
[39]	Topology assaults detection on RPL using semi-automated profiling tool.	● Detection accuracy is high ● Low energy consumption ● Low computation overhead	● High overhead
[40]	Sinkhole attacks are detected using a constraint based specification intrusion detection approach.	● Low overhead ● Minimal energy usage	● Not real-time
[41]	Using a game-theoretic method to identify deceptive attacks in IoT network with honeypots.	● High detection accuracy ● Real-time	● Needs additional resources. ● High converge time

IDS: Intrusion detection system

4. SUPERVISED ML BASED IOT INTRUSION DETECTION

ML enables computer systems to predict events more correctly without being explicitly taught to do so. It is a subset of artificial intelligence (AI). ML algorithms use historical data as input to anticipate new output values. ML algorithms are mainly divided into three categories: reinforcement learning, unsupervised learning, and supervised learning. In this paper, recent researches using supervised ML algorithms in the area of IoT intrusion detection were studied, analyzed and compared. Supervised learning emphasis on discovering patterns while utilizing labeled datasets. In supervised learning, the machine must be fed sample data with different characteristics (expressed as "X") and the right value output of the data (represented as "y"). The dataset is considered "labeled" because the output and feature values are known. Then, the algorithm analyzes data patterns to develop a model that can replicate the same fundamental principles with new data [46].

4.1. Datasets Used for IoT Intrusion Detection

Models for supervised ML are trained and evaluated using datasets. Any IDS's performance ultimately depends on the dataset's quality including whether it can reliably identify assaults or not [47]. Here, six datasets named NSL-KDD, UNSWNB15, CICIDS 2017, Bot-IoT, DS2OS, and IoTID20 are considered and used by researchers to train and test IoT intrusion detection models. Descriptions of the datasets are given below and their characteristics are summarized in Table 5.

- NSL-KDD
  The NSL-KDD dataset is an improved version of the KDD99. It does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records. The number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set [47]. The NSL-KDD dataset has 41 characteristics, classified into three categories: Basic characteristics, content characteristics, and traffic characteristics.

TABLE 4: Comparison of different hybrid IDS techniques

Reference No.	Technique	Strength	Limitations
[42]	Employing a game theoretic approach to identify attackers by using anomaly detection only when a new attack pattern is anticipated and using signature based detection otherwise.	• Detection accuracy is high • Low energy consumption	• High resource consumption • Delay
[43]	The denial of service prevention manager is proposed, which uses aberrant activity detection and matching with attack signatures.	• Real time	• High resource consumption
[44]	Real-time attack detection using knowledgeable, self-adapting expert intrusion detection system.	• High detection accuracy • Real time • Low resource consumption	• High computational overhead
[45]	Attackers can be found by looking for timing irregularities while broadcasting the most recent rank to nearby nodes and using a timestamp.	• Real time • Low overhead • Low delay • High detection accuracy	• High computation overhead • High resource consumption
[27]	Targeting the routing attacks with an IDS with integrated mini-firewall which uses anomaly-based IDS in the intrusion detection and signature-based IDS in the mini-firewall	• Real Time • High availability • Low overhead	• Limited in dynamic network topology • High-resource consumption • Low detection accuracy

IDS: Intrusion detection system

TABLE 5: Dataset characteristics

Dataset	Year	Dataset link (URL)	No. of Instances	No. of Features	Dataset collection performed on IoT environment	Type of dataset
NSLKDD	2009	https://www.unb.ca/cic/datasets/nsl.html	148,519	41	No	Imbalanced
UNSW-NB15	2015	https://research.unsw.edu.au/projects/unsw-nb15-dataset	2,540,044	49	No	Imbalanced
CICIDS2017	2017	https://www.unb.ca/cic/datasets/ids-2017.html	2,830,743	83	No	Imbalanced
BoT- IoT	2019	https://ieee-dataport.org/documents/bot-iot-dataset	73,370,443	29	Yes	Imbalanced
DS2OS	2018	https://www.kaggle.com/datasets/francoisxa/ds2ostraffictraces	409,972	13	Yes	Imbalanced
IoTID20	2020	https://sites.google.com/view/iot-network-intrusion-dataset/home	625,783	83	Yes	Imbalanced

- UNSW-NB15
  The UNSW-NB15 dataset was published in 2015. It was created by establishing the synthetic environment at the UNSW cyber security lab. UNSW-NB15 represents nine major families of attacks by utilizing the IXIA Perfect Storm tool. IXIA tool has provided the capability to generate a modern representative of the real modern normal and the abnormal network traffic in the synthetic environment. There are 49 features and nine types of attack categories known as the analysis, fuzzers, Backdoors, DoS, exploits, reconnaissance, generic, shellcode, and worms [48].
- CICIDS 2017
  The CICIDS 2017 dataset generated in 2017. It includes benign and seven common family of attacks that met real worlds criteria such as DoS, DDoS, brute force, XSS, SQL injection, Infiltration, port scan, and botnet. The dataset is completely labeled with 83 network traffic features extracted and calculated for all benign and attack network flows [49].
- BoT-IoT
  The BoT-IoT dataset was created by designing a testbed network environment in the Research Cyber Range Lab of UNSW Canberra. This dataset consists of legitimate and simulated IoT network traffic along with various types of attacks such as information gathering (probing attacks), denial of service and information theft. It has been labeled with the label features indicating an attack flow, the attacks category and subcategory for possible multiclass classification purposes [50].
- DS2OS
  This dataset includes traces that were recorded using the IoT platform DS2OS. Labeled and unlabeled datasets come in two varieties. The only characteristics in an unlabeled dataset that can be used describe the data objects for unsupervised ML models. In addition, a labeled dataset includes information about each data instance's class and utilized for supervised ML models [51].
- IoTID20
  IoTID20 dataset is used for anomalous activity detection in IoT networks. The testbed for the IoTID20 dataset is a combination of IoT devices and interconnecting structures. The dataset consists of various types of IoT attacks and a large number of flow-based features. The flow-based features can be used to analyze and evaluate a flow-based IDS. The final version of the IoTID20 dataset consists of 83 network features and three label features [52].

4.2. Supervised ML Algorithms Used for IoT Intrusion Detection

For IoT intrusion detection, many supervised ML methods are employed. The list of used algorithms with corresponding descriptions is presented below:

- Logistic regression (LR): It is a probability-based method for predictive analysis. It is a more effective strategy for binary and linear classification issues because it employs the sigmoid function to translate expected values to probabilities between 0 and 1. It is a classification model that is relatively simple to implement and performs extremely well with linearly separable data classes [53].
- Naïve base (NB): Are a group of Bayes' Theorem-based categorization methods. It is a family of algorithms rather than a single method and they all operate under the same guiding principle in which each pair of characteristics is categorized standalone [53].
- Artificial neural networks (ANN): The biological neural network in the human brain served as the model for the widely used ML technology known as (ANN). Each artificial neuron's weight values are sent to the following layer as an output. Feed-forward neural network form of ANN that processes inputs from neurons in the previous layer. Multilayer perception is a significant type of feed forward neural networks (MLP). The most well-known MLP training method that modifies the weights between neurons to reduce error is called the back propagation algorithm. The system can display sluggish convergence and run the danger of a local optimum, but it can rapidly adapt to new data values [54].
- Support Vector Machine (SVM): This algorithm looks for a hyperplane to optimize the distance involving two classes. A learning foundation for upcoming data processing is provided by the categorization. The groups are divided into several configurations by the algorithm through hyperplanes (lines). A learning model that splits up new examples into several categories is produced by SVM. Based on these functions, SVMs are referred to as non-probabilistic, or binary linear classifiers. In situations that use probabilistic classification, SVMs can use methods such as Platt Scaling [53].
- Decision tree (DT) is a tree in which each internal node represents an assessment of an attribute. Each branch represents the result of an assessment and each leaf node denotes the classification outcome. Algorithms such as ID3, CART, C4.5, and C5.0 are frequently used to generate decision trees. By analyzing the samples, a decision tree is obtained and used to correctly classify new data [55].

- Random forest (RF) is a technique used to create a forest of decision trees. This algorithm is frequently used due to its fast operation. Countless decision trees can be used to create a random forest. By averaging the outcomes of each component tree's forecast, this method generates predictions. Random forests exhibit compelling accuracy results and are less likely to overfit the data than a traditional decision tree technique. This method works well while examining plenty of data [53].
- Ensemble Learning (which includes bagging and boosting). The boosting method is a well-known ensemble learning method for improving the performance and accuracy of ML systems. The fundamental idea behind the boosting strategy is the successive addition of models to the ensemble. Weak learners (base learners) are efficiently elevated to strong learners. As a consequence, it aids in reducing variation and bias and raising prediction accuracy. Boosting is an iterative method that alters the findings of an observation's weight depending on the most recent categorization. Adaboost (AB), gradient boosting machines (GBM), and extreme gradient boosting (XGBoost) are examples of boosting techniques. Bagging (also known as bootstrap aggregating). It is one of the earliest and most basic ensemble ML approaches and it works well for issues requiring little in the way of training data. In this approach, a collection of original models with replacement are trained using random subsets of data acquired using the bootstrap sampling method. The individual output models derived from bootstrap samples are combined by majority voting [56].

4.3. Evaluation Metrics

The efficiency of ML algorithms can be measured using metrics such as accuracy, precision, recall, and F1-score [57]. Performance metrics are calculated using different parameters called True positive (TP), False positive (FP), True negative (TN), and False negative (FN). For IDS s, these parameters are described as follow:

TP = The number of cases correctly identified as attack.

FP = The number of cases incorrectly identified as attack.

TN = The number of cases correctly identified as normal.

FN = The number of cases incorrectly identified as normal.

- **Precision** (also called positive predictive value) is the percentage of retrieved occurrences that are relevant. Model performance is considered better if the precision is higher [58]. Precision is computed using (1) [59].

$$\text{Precision} = \frac{\text{True postive}}{\text{True postive} + \text{False postive}} \tag{1}$$

- **Recall** (also known as sensitivity) is the percentage of occurrences that were found to be relevant. It also goes by the name True Positive Rate (TPR) and calculated using (2) [58].

$$\text{Recall} = \frac{\text{True postive}}{\text{True postive} + \text{False negative}} \tag{2}$$

- **Accuracy** is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Model accuracy is calculated using (3) [57].

$$\text{Accuracy} = \frac{\text{True postive} + \text{True negative}}{\text{Total}} \tag{3}$$

- **F1-Score** is the harmonic mean of recall and accuracy [60] which defines a the weighted average of recall and precision and calculated using (4) [57].

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

- **ROC curve** is a receiver operating characteristic curve which shows the performance of a classifier at various thresholds level [57].
- **Area under curve (AUC):** is closely associated with the concept of ROC. It represents the area under the ROC curve. It has been extensively used as a performance measure for classification models in ML. Its values range from 0 to 1. The higher the value, the better the model is [61].

4.4. Analysis and Comparison of Supervised ML Algorithms for IoT Intrusion Detection

In this section, the analysis of the used ML algorithms has been presented and discussed. Researchers used many supervised ML algorithms specifically in classification and they performed well in some cases with very high accuracy. To review researches in the area of intrusion detection using ML in the IoT environment, various recent studies are examined and compared based on the ML algorithms (classifier), datasets, type of classification, and performance of the classifier. The performance of these algorithms depends on various metrics. In this study, the comparison among the algorithms is focused on accuracy metric. Detailed review of 21 papers (published between 2019 and 2022) was analyzed in this section and compared in Table 6.

Mahmudul *et al.* [62] employed the DS2OS dataset with several ML algorithms (LR, SVM, DT, RF, ANN). Accuracy,

TABLE 6: Comparison of the selected supervised ML based IoT IDS

Reference No.	Year	ML algorithm (classifier)	Dataset	Classification type	Classifier accuracy
[62]	2019	LR, SVM, DT, RF, ANN	DS2OS	Multiclass	LR=0.983, SVM=0.982, DT=0.994, RF=0.994, ANN=0.994.
[63]	2019	RF	UNSW-NB15	Binary	RF=99.34
[64]	2019	LR, NB, DT, RF, KNN, SVM	KDD99, NSL-KDD, UNSW-NB15	Binary	Accuracy of the algorithms depend on the used dataset
[65]	2019	For the level-1 model, DT For level 2 model, RF	CICIDS2017, UNSW-15	2 level classification (binary then multiclass)	Both datasets' specificity was 100% for the model, while its precision, recall, and F score were all 100% for the CICIDS2017 dataset and 97% for the UNSW-NB15 dataset
[66]	2019	RF, AB, GBM, XGB, DT (Cart), MLP, extremely randomized trees (ETC)	CIDDS-001, UNSW-NB15, NSL-KDD	Binary	Average accuracy value for 4 datasets using holdout are: RF=94.94, GBM=92.98, XGB=93.15%, AB=90.37, CART=91.98, MLP=82.76, ETC=82.99
[67]	2019	DT, NN, SVM	UNSW-NB15	Multiclass	DT=89.76, NN=86.7و SVM=78.77, Proposed model: 88.92
[68]	2019	NB, QDA, RF, ID3, AB, MLP, KNN	BoT-IoT	Binary.	NB=0.78, QDA=0.88, RF=0.98, ID3=0.99, Adaboost=1.0, MLP=0.84, KNN=0.99
[69]	2019	SVM, LR, D T, KNN, RF	UNSW-NB15, their own dataset	Binary	The accuracy depends on the dataset and the algorithm
[58]	2020	RF, XGB, DT, MLP, GB, ET, LR	UNSW-NB15	Binary	Results with all features: RF=0.9516, XGB=0.9481, DT=0.9387, MLP=0.9371, GB=0.9331, ET=0.9501, LR=0.8984
[53]	2020	KNN, SVM, DT, NB, RF, ANN, LR	Bot-IoT	Binary, multiclass	On binary classification: KNN=0.99, SVM=0.99, DT=1.0, NB=0.99, RF=1.0 ANN=0.99, LR=0.99
[70]	2020	SVM, NB, DT, adaboost	Their own synthetic called (Sensor480)	Binary	SVM=0.9895, NB=0.9789, DT=1.0000, Adaboost=0.9895
[71]	2020	RF	IoTID20 dataset	Binary based on the attack type	The accuracy result depends on the attack type
[72]	2021	SVM	NSL-KDD, UNSW-NB15.	Binary, multiclass	The accuracy depends on the dataset, the type of classification and number of features
[55]	2021	RF, SVM, ANN	UNSW-NB15.	Binary, multiclass	All features: RF with Binary=98.67, Multi-class=97.37, SVM in Binary=97.69, Multiclass=95.67, ANN in Binary=94.78, multiclass=91.67
[73]	2021	LR, SVM, DT, ANN	IoTID20, BoT-IoT	Multiclass	The results are based on the dataset and the categories of attacks
[74]	2021	SLFN	IoTID20	Binary	The proposed model=0.9351
[75]	2021	SVM, GBDT, RF	NSL KDD	Binary	SVM=32.38, GBDT=78.01, RF=85.34
[76]	2021	B-stacking	CICIDS2017, NSL-KDD	Multiclass	Accuracy for CICIDS2017 is 99.11% Accuracy for NSL-KDD approximately is 98.5%
[77]	2022	DT, RF, GBM	IoT2020	Binary	DT=0.978305, RF=0.978443, GBM=0.9636
[78]	2022	Shallow neural networks (SNN), bagging trees (BT), DT, SVM, KNN	IoTID20	Binary, multiclass	For binary classification all models achieved 100% For multiclass: SNN=100%, DT=99.9%, BT=99.9%, SVM=99,8%, KNN=99.4%
[79]	2022	ANN, DT (C4.5), Bagging, KNN, Ensemble	IoTID20, NSL-KDD	Binary, multiclass	Accuracy depends on feature selection approaches, datasets, and attack type for multiclass classification

precision, recall, f1 score, and area under the receiver operating characteristic curve are the assessment measures used to compare performance. The measurements show that RF performs comparably higher performance, and the system acquired excellent accuracy (Ibrahim *et al.* [63]). An intelligent anomaly detection system called Anomaly Detection IoT (AD-IoT) which used the UNSW-NB15 dataset and RF to identify binary labeled categorization had been proposed. The results demonstrated that the AD-IoT could successfully produce the best classification accuracy while minimizing the false positive rate. Samir *et al.* in [64] used the datasets KDD99, NSL-KDD, and UNSW-NB15 to assess number of ML models. The KKN and LR algorithms produced the best results on the UNSW-NB15 dataset while the NB algorithm produced the worst results. On the NSL-KDD dataset, the DT classifier outperformed the others in terms of various metrics while on the KDD99 dataset, SVM and MLP produced a low false positive rate in comparison to other algorithms. The findings of this study showed that the DT and KNN algorithms outperformed the other algorithms. However, the KNN required more time to categorize data than the DT. Imtiaz and Qusay [65] conducted a two-level framework experiment for IoT intrusion detection. To determine the category of the anomaly, they chose a DT classifier for the level-1 model which categorized the network flow as normal or anomalous and forwarded the network flow to the level-2 model. RF was used as a level-2 model for multiclass categorization. Abhishek and Virender [66] employed both ensemble and single classifiers, two different types of classification techniques. The selection of the aforementioned classification algorithms was primarily influenced by the huge number of input characteristics that are vulnerable to overfitting. As a result, random search was used to determine the best input parameters for RF, AB, XGB, and GBM. In terms of precision, RF beats other classifiers. However, AB performs the worst of all the classifiers. Using Friedman test statistics and 10-fold validation, the results showed that the classifiers' performances are considerably varied. Following that, the average time required by several classifiers to categorize a single case, CART classifies instances of CIDDS-001, UNSW-NB15, KDDTrain+, and KDDTest+ faster than other classifiers. Vikash *et al.* [67] proposed (UIDS) an IDS using UNSW-NB15 dataset. Network traffic accuracy and assault detection rate were improved by the suggested approach. In addition, it examined data using several ML techniques (C5, neural network, SVM, and UIDS model) and came to the conclusion that UIDS compared favorably to other ML techniques. Analysis showed that the false alarm rate (FAR) of the UNSW-NB15 dataset was reduced with only 13 characteristics. Jadel and Khalid [68] tested seven ML algorithms. All the algorithms, except the Naive Bayes (NB) and Quadratic algorithm (QDA), achieved highest success in detecting almost all attack types. It can be seen that Adaboost was the best performance algorithm, followed by KNN and ID3. ID3 is noticeably faster than KNN. The accuracy of the algorithms depends on the entire dataset with the seven best features obtained in the feature selection step. Aritro *et al.* [69] analyzed the role of a set of chosen ML techniques for IoT intrusion detection based on dataset/flows two layers: Application layer (host based) and network layer (network based). For the application layer dataset, they created their own dataset from the IoT environment while for network layer they used UNSW-NB15 dataset. According to the results for both datasets, RF was the best algorithm in terms of accuracy and LR was the fastest in terms of speed. Mohammad [58] used different algorithms. The classifiers random forest (RF) and extra trees (ET) performed better than the others, and RF is the best of the two. Only 14 features were chosen by RF utilizing features selection, but the performance results were remarkably similar to those achieved with all features. In addition, compared to the others, the LR classifier had the lowest accuracy. Andrew *et al.* [53] employed different methods; nevertheless, the findings show that RF performed better with the non-weighted dataset regarding precision and accuracy in non-weighted dataset. However, ANN performed more accurately in binary classification using weighted dataset. KNN and ANN performed extremely well in multi-classification for weighted and non-weighted datasets, respectively. The findings made it clear that ANN accurately predicted the kind of attack. K. V. V. N. L *et al.* [70] tested four ML techniques on IoT traffic in order to distinguish between genuine and attack traffic. Using decision trees, all of the analyzed data may be precisely categorized into the correct classes. Decision trees also had the greatest accuracy compared to the other classifiers. Pascal *et al.* [71] suggested a new anomaly-based detection using hybrid feature selection for IoT networks using IoTID20 dataset. The relevant features were fed to the RF algorithm. Based on the attack category, the network traffic is classified as normal and attack category as DoS, Scan, or MITM. Nsikak *et al.* [72] tested SVM with dataset NSL-KDD and UNSW-NB15 datasets. The results using different numbers of features for both datasets were varied. The classification accuracy using binary classification was greater than multi-class according to the evaluation results. Muhammad *et al.* [55], the UNSW-NB15 dataset had been subjected to supervise ML including RF, SVM, and ANN.

The application of RF using mean imputation produced the greatest accuracy in binary classification. Overall, there were not many differences in accuracy across the different imputation strategies. By using RF on a regression-imputed dataset, the greatest accuracy in multi-class classification was also attained. In addition, as compared to other cutting-edge supervised ML-based techniques, RF achieved greater accuracy with less training time for clustered based classification. Khalid *et al.* [73] for classification objectives, the performance of four ML methods was assessed. The Bot-IoT dataset and the IoTID20 dataset were both utilized in the study, 5% of Bot-IoT dataset was selected with a full set of features, while the second dataset was fully selected in the experiment. The accuracy results were based on the dataset and the categories of attacks. Raneem *et al.* [74] developed an intrusion detection method using a single layer forward neural network (SLFN) classifier with IoTID20. The results showed that the SLFN classification approach outperformed other classification algorithms. Maryam *et al.* [75] proposed that three ML algorithms RF, GDBT, and SVM were applied to the NSL-KDD dataset using binary classification. The results showed that the RF obtained the highest accuracy on the fog layer while SVM obtained lowest accuracy. Souradipst *et al.* [76] proposed B-Stacking approach as an intrusion detection model to detect cyber-attacks and anomalies in IoT networks. B-Stacking is based on a combination of two ensemble algorithms; boosting and stacking. It chose KNN, RF, and XGBoost as the level-0 weak learners. XGboost is also used as the level-1 learner. The experimental results on two popular datasets showed that the model had a high detection rate and a low false alarm rate. Most importantly, the proposed model is lightweight and can be deployed on IoT nodes with limited power and storage capabilities. Jingyi *et al.* [77] used DT, RF, and GBM ML algorithms with a dataset generated from the IoTID20 dataset known as IoT2020 dataset. According to the results, the DT algorithm performed more accurately than the other algorithms, but RF had better AUC score. Abdulaziz *et al.* [78] proposed an anomaly intrusion detection in an IoT system. Five supervised ML models were implemented to characterize their performance in detecting and classifying network activities with feature engineering and data preprocessing framework. Based on experimental evaluation, the accuracy 100% recorded for the detection phase that distinguishes the normal and anomaly network activities. While for classifying network traffic into five attack categories, the implemented models of achieved 99.4-99.9%. Khalid *et al.* [79] proposed and implemented an IoT anomaly-based IDS based on novel feature selection and extraction approach. The model framework was trained and tested on IoTID20 and NSL-KDD datasets using four ML algorithms. The system scored a maximum detection accuracy of 99.98% for the proposed ML ensemble-based hybrid feature selection approach.

From the literature, it is observed that there are extensive efforts on developing IDS s for IoT. Several researchers have assessed the effectiveness of their systems using common datasets like NSL-KDD, UNSW-NB15, and CICIDS2017. These datasets were not used captured traffic from IoT environment. Hence, an extensive work should be conducted using recent datasets such as IoTID20 which consists of IoT network traffic features. The state of the art also shows that some models perform well, particularly tree-based algorithms such as boosting, random forest and decision trees. ML algorithms' performance outcomes vary depending on the used dataset, features, and classification category.

5. CONCLUSION

One of the most important technological progresses over the past decade was the widespread adoption of IoT devices across industries and societies. With the development of IoT, several obstacles have been raised. One of these obstacles is IoT security which cannot be disregarded. IoT networks are vulnerable to a variety of threats. Although the IoT network is protected by encryption and authentication, cyber attacks are still possible. Therefore, using IoT IDS is important and necessary. This paper conducted an in-depth comprehensive analysis and comparison of various recent researches which used different techniques, datasets, ML algorithms and their performance for detecting IoT intrusions. Based upon the analysis, the recent IoT dataset for intrusion detection is identified which is IoTID20 dataset. Furthermore, the ML algorithms that outperformed in most researches are tree-based algorithms such as DT, RF, and boosting algorithms. Many points were observed and needed further study like using and collecting real IoT intrusion detection datasets for training and testing ML models, real time, and lightweight IDSs are required that need less detection time and resources consumption. All these factors should be taken into account while developing new IoT IDSs. In addition, further study should be conducted to address recent IoT threats, and the need to identify the best IDS placement techniques that improve IoT security while lowering the risk of cyber attacks.

REFERENCES

[1] S. Chen, H. Xu, D. Liu, B. Hu and H. Wang. "A vision of IoT: Applications, challenges, and opportunities with china perspective."

IEEE Internet of Things Journal, vol. 1, no. 4, pp. 349-359, 2014.

[2] S. Li, L. D. Xu and S. Zhao. "The internet of things: A survey". *Information Systems Frontiers*, vol. 17, no. 2, pp. 243-259, 2015.

[3] T. Sherasiya and H. Upadhyay. "Intrusion detection system for internet of things". International Journal of Advance Research and Innovative Ideas in Education, vol. 2, no. 3, pp. 2244-2249, 2016.

[4] M. M. Patel and A. Aggarwal. "*Security Attacks in Wireless Sensor Networks: A Survey*". In: 2013 International Conference on Intelligent Systems and Signal Processing (ISSP). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 329-333, 2013.

[5] S. N. Kumar. "Review on network security and cryptography". *International Transaction of Electrical and Computer Engineers System*, vol. 3, no. 1, pp. 1-11, 2015.

[6] R. S. M. Joshitta, L. Arockiam. "Security in IoT environment: A survey". *International Journal of Information Technology and Mechanical Engineering*, vol. 2, no. 7, pp. 1-8, 2016.

[7] M. M. Hossain, M. Fotouhi and R. Hasan. "*Towards an Analysis of Security Issues, Challenges, and Open Problems in the Internet of Things*". In: 2015 IEEE World Congress on Services. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 21-28, 2015.

[8] A. Khraisat and A. Alazab. "A critical review of intrusion detection systems in the internet of things: Techniques, deployment strategy, validation strategy, attacks, public datasets and challenges". *Cybersecurity*, vol. 4, no. 1, pp. 1-27, 2021.

[9] N. Mishra and S. Pandya. "Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review". *IEEE Access*, vol. 9, pp. 59353-59377, 2021.

[10] L. Atzori, A. Iera and G. Morabito. "The internet of things: A survey," *Journal of Computer Network*, vol. 54, no. 15, pp. 2787-2805, 2010.

[11] S. Andreev and Y. Koucheryavy. "*Internet of things, smart spaces, and next generation networking*". vol. 7469. In: Lecture Notes in Computer Science. Springer, Berlin, Germany, p. 464, 2012.

[12] S. J. Kumar and D. R. Patel. "A survey on internet of things: Security and privacy issues". *International Journal of Computer Applications*, vol. 90, no. 11, pp. 20-26, 2014.

[13] J. Du and S. Chao. "*A Study of Information Security for M2M of IOT*". In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). Vol. 3. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. V3-576-V3-579, 2010.

[14] B. Schneier. *Secrets and Lies: Digital Security in a Networked World*. John Wiley and Sons, Hoboken, New Jersey, 2015.

[15] J. M. Kizza. *Guide to computer network security*. Springer, Berlin, Germany, 2013.

[16] M. Taneja. "*An analytics framework to detect compromised IoT devices using mobility behavior*". In: 2013 International Conference on ICT Convergence (ICTC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 38-43, 2013.

[17] G. M. Koien and V. A. Oleshchuk. "*Aspects of Personal Privacy in Communications: Problems, Technology and Solutions*". River Publishers, Denmark, 2013.

[18] N. R. Prasad. "*Threat Model Framework and Methodology for Personal Networks* (*PNs*)". In: 2007 2nd International Conference on Communication Systems Software and Middleware. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2007.

[19] S. O. Amin, M. S. Siddiqui, C. S. Hong, and J. Choe. "*A novel coding scheme to implement signature based IDS in IP based Sensor Networks*". In: 2009 IFIP/IEEE International Symposium on Integrated Network Management-workshops. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 269-274, 2009.

[20] J. Deogirikar and A. Vidhate. "*Security Attacks in IoT: A Survey*". In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 32-37, 2017.

[21] S. Ansari, S. Rajeev and H. S. Chandrashekar. "Packet sniffing: A brief introduction". *IEEE Potentials*, vol. 21, no. 5, pp. 17-19, 2003.

[22] L. Liang, K. Zheng, Q. Sheng and X. Huang. "*A Denial of Service Attack Method for an IoT System*". In: 2016 8th International Conference on Information Technology in Medicine and Education (ITME). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 360-364, 2016.

[23] C. Wilson. "*Botnets, Cybercrime, and Cyberterrorism: Vulnerabilities and Policy Issues for Congress*". Library of Congress, Congressional Research Service, Washington, DC, 2008.

[24] K. Tsiknas, D. Taketzis, K. Demertzis, and C. Skianis. "Cyber threats to industrial IoT: A survey on attacks and countermeasures". *IoT*, vol. 2, no. 1, pp. 163-186, 2021.

[25] N. Chakraborty and B. Research. "Intrusion detection system and intrusion prevention system: A comparative study". *International Journal of Computing and Business Research*, vol. 4, no. 2, pp. 1-8, 2013.

[26] N. Das, T. Sarkar. "Survey on host and network-based intrusion detection system". *International Journal of Advanced Networking and Applications*, vol. 6, no. 2, p. 2266, 2014.

[27] S. Raza, L. Wallgren and T. Voigt. "SVELTE: Real-time intrusion detection in the internet of things". *Ad Hoc Networks*, vol. 11, no. 8, pp. 2661-2674, 2013.

[28] P. Y. Chen, S. M. Cheng and K. C. Chen. "Information fusion to defend intentional attack in internet of things". *IEEE Internet of Things Journal*, vol. 1, no. 4, pp. 337-348, 2014.

[29] P. Pongle and G. Chavan. "Real time intrusion and wormhole attack detection in internet of things". *International Journal of Computer Applications*, vol. 121, no. 9, pp. 1-9. 2015.

[30] C. Cervantes, D. Poplade, M. Nogueira and A. Santos. "*Detection of Sinkhole Attacks for Supporting Secure Routing on 6LoWPAN for Internet of Things*". In: 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 606-611, 2015.

[31] D. H. Summerville, K. M. Zach and Y. Chen. "*Ultra-lightweight Deep Packet Anomaly Detection for Internet of Things devices*". In: 2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-8, 2015.

[32] V. Eliseev and A. Gurina. "*Algorithms for Network Server Anomaly Behavior Detection without Traffic Content Inspection*". In: Proceedings of the 9th International Conference on Security of Information and Networks. Association for Computing Machinery, New York, pp. 67-71, 2016.

[33] S. O. Amin, M. S. Siddiqui, C. S. Hong and S. Lee. "Implementing signature based IDS in IP-based sensor networks with the help of signature-codes". *IEICE Transactions on Communications*, vol. 93,

no. 2, pp. 389-391, 2010.

[34] D. Oh, D. Kim and W. W. Ro. “A malicious pattern detection engine for embedded security systems in the internet of things”. *Sensors*, vol. 14, no. 12, pp. 24188-24211, 2014.

[35] H. Sun, X. Wang, R. Buyya and J. Su. “CloudEyes: Cloud-based malware detection with reversible sketch for resource-constrained internet of things (IoT) devices”. *Journal of Software Practice and Experience*, vol. 47, no. 3, pp. 421-441, 2017.

[36] L. Santos, C. Rabadao and R. Gonçalves. “*Intrusion Detection Systems in Internet of Things: A Literature Review*”. In: 2018 13th Iberian Conference on Information Systems and Technologies (CISTI). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-7, 2018.

[37] F. Ahmed, Y. B. Ko. “Mitigation of black hole attacks in routing protocol for low power and lossy networks”. *Security and Communication Network*s, vol. 9, no. 18, pp. 5143-5154, 2016.

[38] Y. Xia, H. Lin and L. Xu, “*An AGV Mechanism Based Secure Routing Protocol for Internet of Things*”. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 662-666, 2015.

[39] A. Le, J. Loo, K. K. Chai and M. Aiash. “A specification-based IDS for detecting attacks on RPL-based network topology”. *Information*, vol. 7, no. 2, p. 25, 2016.

[40] M. Surendar and A. Umamakeswari. “*InDReS: An Intrusion Detection and Response System for Internet of Things with 6LoWPAN*.” In: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1903-1908, 2016.

[41] Q. D. La, T. Q. S. Quek, J. Lee, S. Jin and H. Zhu. “Deceptive attack and defense game in honeypot-enabled networks for the internet of things”. *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1025-1035, 2016.

[42] H. Sedjelmaci, S. M. Senouci and M. Al-Bahri. “*A Lightweight Anomaly Detection Technique for Low-resource IoT Devices: A Game-theoretic Methodology*”. In: 2016 IEEE International Conference on Communications (ICC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6 2016.

[43] P. Kasinathan, C. Pastrone, M. A. Spirito and M. Vinkovits. “*Denial-of-Service detection in 6LoWPAN based Internet of Things*.” In: 2013 IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 600-607, 2013.

[44] D. Midi, A. Rullo, A. Mudgerikar, and E. Bertino. “*Kalis-a System for Knowledge-driven Adaptable Intrusion Detection for the Internet of Things*”. In: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 656-666, 2017.

[45] T. Matsunaga, K. Toyoda and I. Sasase. “Low false alarm attackers detection in RPL by considering timing inconstancy between the rank measurements”. *IEICE Communications Express*, vol. 4, no. 2, pp. 44-49, 2015.

[46] M. Praveena and V. Jaiganesh. “A literature review on supervised machine learning algorithms and boosting process”. *International Journal of Computer Applications*, vol. 169, no. 8, pp. 32-35, 2017.

[47] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani. “*A Detailed Analysis of the KDD CUP 99 Data Set*”. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2009.

[48] N. Moustafa and J. Slay. “*UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems* (*UNSW-NB15 Network Data Set*)”. In: 2015 Military Communications and Information Systems Conference (MilCIS). IEEE. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2015.

[49] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani. “*Toward generating a new intrusion detection dataset and intrusion traffic characterization*”.In: The International Conference on Information Systems Security and Privacy. vol. 1, pp. 108-116, 2018.

[50] N. Koroniotis, N. Moustafa, E. Sitnikova and B. Turnbull. “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset”. *Future Generation Computer Systems*, vol. 100, pp. 779-796, 2019.

[51] F. X. Aubet. “Machine Learning-Based Adaptive Anomaly Detection in Smart Spaces”. B.Sc. Thesis, Department of Informatics, Technische Universität München, Germany, 2018.

[52] I. Ullah and Q. H. Mahmoud. “*A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks*”. In: Canadian Conference on Artificial Intelligence. Springer, Berlin, Germany, pp. 508-520, 2020.

[53] A. Churcher, R. Ullah, J. Ahmad, S. U. Rehman, F. Masood, M. Gogate, F. Alqahtani, B. Nour and W. J. Buchanan. “An experimental analysis of attack classification using machine learning in IoT networks”. *Sensors*, vol. 21, no. 2, p. 446, 2021.

[54] R. Olivas. “Decision Trees,” Rafael Olivas, San Francisco, 2007.

[55] M. Ahmad, Q. Riaz, M. Zeeshan, H. Tahir, S. A. Haider, M. S. Khan. “Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using UNSW-NB15 data-set”. *Journal on Wireless Communications and Networking*, vol. 2021, no. 1, pp. 1-23, 2021.

[56] J. Dou, A. P. Yunus, D. T. Bui, A. Merghadi, M. Sahana, Z. Zhu, C. W. Chen, Z. Han, B. T. Pham. “Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan”. *Landslide*, vol. 17, no. 3, pp. 641-658, 2020.

[57] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan. “Performance analysis of machine learning algorithms in intrusion detection system: A review”. *Procedia Computer Science*, vol. 171, pp. 1251-1260, 2020.

[58] M. Shorfuzzaman. “*Detection of Cyber Attacks in IoT using Tree-based Ensemble and Feedforward Neural Network*”. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 2601-2606, 2020.

[59] D. L. Streiner and G. R. Norman. “Precision” and “accuracy”: Two terms that are neither”. *Journal of Clinical Epidemiology*, vol. 59, no. 4, pp. 327-330, 2006.

[60] D. Chicco and G. Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[61] W. Ma and M. A. Lejeune. “A distributionally robust area under curve maximization model”. *Operations Research Letters*, vol. 48, no. 4, pp. 460-466, 2020.

[62] M. Hasan, M. M. Islam, M. I. I. Zarif and M. M. A. Hashem. “Attack and anomaly detection in IoT sensors in IoT sites using machine

learning approaches". *Internet of Things*, vol. 7, p. 100059, 2019.

[63] I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy and H. Ming. "*Ad-iot: Anomaly Detection of IOT Cyberattacks in Smart City Using Machine Learning*". In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 0305-0310, 2019.

[64] S. Fenanir, F. Semchedine and A. Baadache. "A machine learning-based lightweight intrusion detection system for the internet of things". *Revue D Intelligence Artificielle*, vol. 33, no. 3, pp. 203-211, 2019.

[65] I. Ullah and Q. H. Mahmoud. "*A Two-level Hybrid Model for Anomalous Activity Detection in IoT Networks*". In: 2019 16th IEEE Annual Consumer Communications and Networking Conference (CCNC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2019.

[66] A. Verma and V. Ranga. "Machine learning based intrusion detection systems for IoT applications". *Wireless Personal Communications*, vol. 111, no. 4, pp. 2287-2310, 2020.

[67] V. Kumar, A. K. Das, and D. Sinha. "UIDS: A unified intrusion detection system for IoT environment". *Evolutionary Intelligence*, vol. 14, no. 1, pp. 47-59, 2021.

[68] J. Alsamiri and K. Alsubhi. "Internet of things cyber attacks detection using machine learning". *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 628-634, 2019.

[69] A. R. Arko, S. H. Khan, A. Preety and M. H. Biswas. "*Anomaly Detection In IoT using Machine Learning Algorithms*". Brac University, Bangladesh, 2019.

[70] K. V. V. N. L. S. Kiran, R. N. K. Devisetty, N. P. Kalyan, K. Mukundini, and R. Karthi. "Building a intrusion detection system for IoT environment using machine learning techniques". *Procedia Computer Science*, vol. 171, pp. 2372-2379, 2020.

[71] P. Maniriho, E. Niyigaba, Z. Bizimana, V. Twiringiyimana, L. J. Mahoro and T. Ahmad. "*Anomaly-based Intrusion Detection Approach for IOT Networks Using Machine Learning*". In: 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 303-308, 2020.

[72] N. P. Owoh, M. M. Singh, Z. F. Zaaba, and Applications. "A hybrid intrusion detection model for identification of threats in internet of things environment".*International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, pp. 689-697, 2021.

[73] K. Albulayhi, A. A. Smadi, F. T. Sheldon and R. K. Abercrombie. "IoT intrusion detection taxonomy, reference architecture, and analyses". *Sensors*, vol. 21, no. 19, p. 6432, 2021.

[74] R. Qaddoura, A. M. Al-Zoubi, H. Faris and I. Almomani. "A multi-layer classification approach for intrusion detection in iot networks based on deep learning". *Sensors*, vol. 21, no. 9, p. 2987, 2021.

[75] M. Anwer, S. M. Khan, M. U. Farooq and W. Nazir. "Attack detection in IoT using Machine Learning". *Engineering Technology and Applied Science Research*, vol. 11, no. 3, pp. 7273-7278, 2021.

[76] S. Roy, J. Li, B. J. Choi and Y. Bai. "A lightweight supervised intrusion detection mechanism for IoT networks". *Future Generation Computer Systems*, vol. 127, pp. 276-285, 2022.

[77] J. Su, S. He and Y. Wu. "Features selection and prediction for IoT attacks". *High Confidence Computing*, vol. 2, no. 2, p. 100047, 2022.

[78] A. A. Alsulami, Q. Abu Al-Haija, A. Tayeb, and A. Alqahtani, "An Intrusion Detection and Classification System for IoT Traffic with Improved Data Engineering". *Applied Sciences*, vol. 12, no. 23, p. 12336, 2022.

[79] K. Albulayhi, Q. A. Al-Haija, S. A. Alsuhibany, A. A. Jillepalli, M. Ashrafuzzaman and F. T. Sheldon. "IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method". *Applied Sciences*, vol. 12, no. 10, p. 5015, 2022.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Liver Fluke Species Identification Isolated From Humans and Animals Using PCR-RFLP and DNA Sequencing

Vilya Sh. Othman[1*], Abdullah A. Hama[1,2], Dana T. Garib[3] and Rostam H. Zorab[4]

[1]*Medical Laboratory Department, College of Health and Medical Technology, Sulaimani Polytechnic University, Kurdistan Region, Iraq,* [2]*MLS, College of Health Science, University of Human Development, Kurdistan Region, Iraq,* [3]*Gastroenterology and Hepatology Center,General Health Directorate of Sulaimani,Kurdistan Region, Iraq,* [4]*Slemani Veterinary Directorate, Kurdistan Region,Iraq*

ABSTRACT

Fasciola species are a member of flatworms belonging to the trematodes (flukes), commonly known as liver fluke, they are extremely pathogenic parasites that affect the liver of humans and animals, nowadays, most laboratories and research facilities use molecular-based techniques for identifying and describing *Fasciola* species. The molecular diagnostic markers such as polymerase chain reaction (PCR), restriction fragment length polymorphism (RFLP), and PCR-RFLP methods are accurate and more specific than the immunological and microscopical methods. The identification of the species of liver flukes will give a new clue for the treatment and control of fascioliasis. The aim, of this study, is to find the molecular characterization of *Fasciola* spp. isolated from humans and animals in Sulaimani city. The flukes were isolated from humans using endoscopic techniques and from slaughtered livestock at the new slaughterhouse of Sulaimani, 48 liver flukes were collected from different hosts; human (n = 3), cattle (n = 20), sheep (n = 20), and goats (n = 5) from October 2021 to April 2022. The uinversal primers ribosomal Deoxy Ribo Nuclic Acid (rDNA) were used, then the PCR products were subjected to restriction fragment polymorphism (RFLP) assay and The PCR Product was digested with restriction enzymes DraII, also the DNA sequencing was used for the PCR product of the primer Cytochrome Oxidase subuint 1 (*COX1*). The results of the PCR-RFLP of the 28s rDNA show the genetic polymorphisms among the flukes and two patterns of RFLP were observed *F. hepatica*, and *F. gigantica*, also the sequence analysis of the partial gene of the COX1 showed the isolated flukes belonged to *F. hepatica* and *F. gigantica* with some genetic variation, and the result of the sequences was deposited in the Gene Bank under the following Accession numbers; *F. gigantica* (OP718780 and OP718781) and *F. hepatica* (OP718782, OP718783, and OP718784). The present study concludes that *F. hepatica* and *F. gigantica* are both responsible for human and animal Fasciolasis in Kurdistan-Iraq, Therefore, RFLP techniques and DNA sequencing are a reliable, and differential method for species and genotyping identification of liver fluke.

Index Terms: *Fasciola hepatica*, *Fasciola gigantica*, Polymerase chain reaction, Restriction fragment length polymorphism, Liver fluke

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp66-70

E-ISSN: 2521-4217

P-ISSN: 2521-4209

1. INTRODUCTION

The Fasciola species is an extremely pathogenic parasite that affects humans and animals and this disease has a dual impact on human health and economic losses [1], [2]. The parasite habitat is the livers and bile ducts of humans and animals.

Corresponding author's e-mail: Vilya Sh. Othman, Department of Medical Laboratory, College of Health Science, Sulaimani Polytechnic University, Kurdistan Region, Iraq. email: viliashwan22@gmail.com

Received: 15-12-2022 Accepted: 19-02-2023 Published: 08-03-2023

In livestock, it causes numerous financial losses, such as a decrease in milk, meat, and wool production [3]. This fluke directly affects livestock productivity through host mortality and partial or complete liver condemnation. Furthermore, it has an indirect effect on the host growth and feeding, which negatively reflect on the white and milk yield [4]. The effect of animal Fasciolasis on weight will be varied depending on the age group and poorer carcass quality indicators [5]. The human Fascioilasis which is known as a neglected disease is occurring in the area where the peoples eat raw water plants that may be contaminated with metacercaria [6]. Due to the epidemiological traits of *Fasciola hepatica* and *Fasciola gigantica* infections in humans, accurate and precise diagnoses are crucial for early uncomplicated treatment. The species of liver fluke have differential characteristics and aspects, also the different type of lymnaeid snails serve as the intermediate host of the liver flukes [7]. The clinical, pathological, and immunological diagnosis approaches are not efficient methods to differentiate between the species of liver fluke [7], [8]. Furthermore, the morphometric methods faced many limitations, the differentiation of *F. hepatica* and *F. gigantica* is currently based on implementing numerous molecular methods using various DNA markers [7]. Various molecular diagnostic tools, including polymerase chain reaction (PCR), PCR-restriction fragment length polymorphism (RFLP), and *rDNA* sequencing, have been used to find the genetic variation and species identification of *Fasciola* spp. [9], [10], [11], [12]. Furthermore, due to the shortcomings of the morphometric approach, many molecular strategies have been applied for genotyping and classifying liver flukes throughout the world. These strategies use various molecular markers [13]. A quick and straightforward PCR-RFLP experiment using the common restriction enzyme Dra II is presented to differentiate between species of liver flukes. Based on the 28S, rDNA gene sequence recently discovered the polymorphism among liver fluke populations [9], [14]. The rDNA region has been applied to establish the genetic polymorphism of *Fasciola* spp. in Spain. This revealed that heterozygous specimens had nucleotide variants [15]. In a new study, a genetically varied type of liver fluke was discovered, named *Fasciola intermediate*, it is morphologically similar to *F. gigantica* and *F. hepatica* [16]. The present study is an attempt to find the species of the liver flukes isolated from humans and animals using PCR-RFLP and DNA sequencing.

2. MATERIALS AND METHODS

2.1. Sample and Data Collection

The 48 liver flukes were collected from different hosts; humans (*n* = 3), cattle (*n* = 20), sheep (*n* = 20), and goats (*n* = 5). The data on human fascioliasis and the flukes was taken from the Kurdistan Center for Gastroenterology and Hepatology in Sulaimani. Furthermore, the animal data were collected in the new Sulaimani slaughterhouse from November 2021 to April 2022. The flukes were preserved in 70% ethanol, then transported to the laboratory of Sulaimani Polytechnic University and the genetic laboratory of the University of Human Development in Sulaimani Kurdistan-Iraq.

2.2. DNA Extraction and PCR

The DNA was extracted from the tissue of the isolated worm (whole body of the worm), the liver flukes were ground individually in an Eppendorf tube using micro-pistil, and the commercial DNA extraction kit (ADD prep, Tissue Genomic DNA extraction kit) was used. The extracted DNAs were stored at −20°C.

In the present study, two molecular approaches were followed RFLP and DNA sequencing, and two sets of primers were used Table 1. The 30 μL PCR reactions prepared to amplify the target region of the liver fluke DNA contained 2 μL of the target DNA (sample), 15 μL of master mix 2× (Amplicon, Skovlunde, Denmark), 1 μL of forwarding primer, 1 μL of reverse primer, and 11 μL deionized distilled water (ddH$_2$O), the conventional PCR was used (Thermocycler BIO-RAD iQ TM), the program of the thermo-cycler was set for both primers as follow; initial denaturation 5 min at 95°C followed by 35 cycles at 95°C for 50 s, 55°C for 40 s, and 72°C for 1 min with a final extension of 7 min at 72°C.

2.3. DNA Digestion by Restriction Enzyme

The detection of genetic variation and species identification of the liver fluke was achieved by a restriction fragment length polymorphism (RFLP) assay. The Dra II restriction enzyme (NEB, UK) is used for the digestion of the PCR product of the ribosomal 28s rDNA primer as described

Table 1: The nucleotide sequences of primers used in the present study

Primer name	Sequences of the primers	References
28S rDNA	F 5’-ACG TGA TTA CCC GCT GAA CT–3’ R 5’ –CTG AGA AAG TGC ACT GAC AAG–3’	[14]
FCOX1	F 5’–AAA TGC TTT GAG TGC TTG GTTG–3’ R 5’–ATG AGC AAC CAC AAA CCA CG–3’	[15]

previously [14], The RFLP reactions contained 4 µL PCR product, 1 µL restriction enzymes, 5 µL buffers, and 30 µL ddH$_2$O. Then, the reaction mixture was incubated at 37°C for 15 min followed by heat-inactivated at 65°C for 20 min. While the PCR product of the *FCOX1* was subjected to DNA sequencing and the result of the DNA sequences was assembled, analyzed, aligned, and blasted in NCBI with the previously registered, the sequence editing software (Bioedit software) was used.

2.4. Gel Electrophoresis

The 5 µL of PCR product was run on 1.5% agarose gel, stained with 5 µL of safe Dye (ADDBIO INC), the gel electrophoresis was run at 84 V for 60 min and visualized under UV (Biobase- China) automatic gel imaging and analysis system in all amplifications. Furthermore, 5 µL of the digested PCR product was run on 2% agarose gels at 84 V for 90 min, 5 µL of safe Dye was added to the gel for DNA staining and visualized under a UV illuminator, also the DNA sequences were submitted to the Gene bank (NCBI) and got the accession numbers.

3. RESULTS

Out of 48 liver flukes isolated from humans and animals, all PCR runs successfully amplified the suspected band (618 bp) for 28S rDNA primer and (836 bp) for the *FCOX1* primer (Figs. 1 and 2). Furthermore, for genetic characterization and species identification of liver flukes, DNA sequencing for *FCOX1* and RFLP were done.

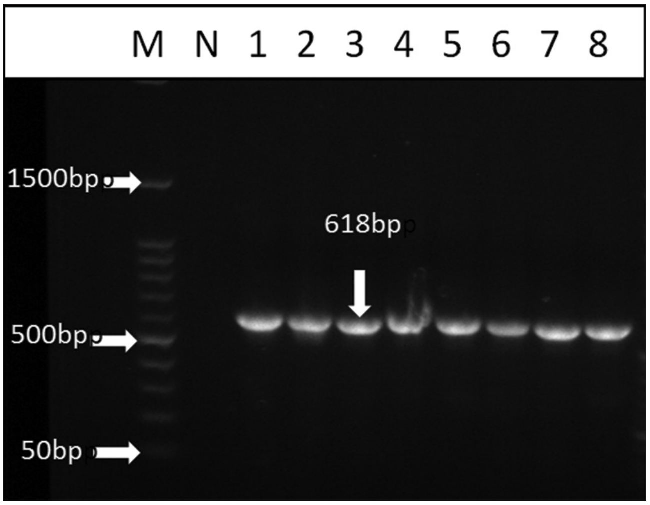


Fig. 1. Gel electrophoresis of the PCR product of the primer 28S rDNA. M= DNA ladder (50 bp), N = Negative control, 1–8 = The samples.

The result of RFLP analysis of the PCR product of the 28S rDNA primer (618 bp), which is digested by restriction enzyme (DraII) showed two patterns, the first one has two bands of (529 bp and 89 bp), while the other pattern is not digested and the 618bp band remains. Table 2 and Fig. 3, the first pattern belongs to *F. hepatica*, and the second is *F. gigantica* [15]. The DNA sequencing of the PCR product of the FCOX1 primers showed the genetic variation, and the result of the nucleotide analysis indicates the presence of *F. hepatica* and *F. gigantica* in the Kurdistan-Iraq, the results were deposited in the Genbank the following Accession numbers; (OP718780 and OP718781) and (OP718782, OP718783, and OP718784).

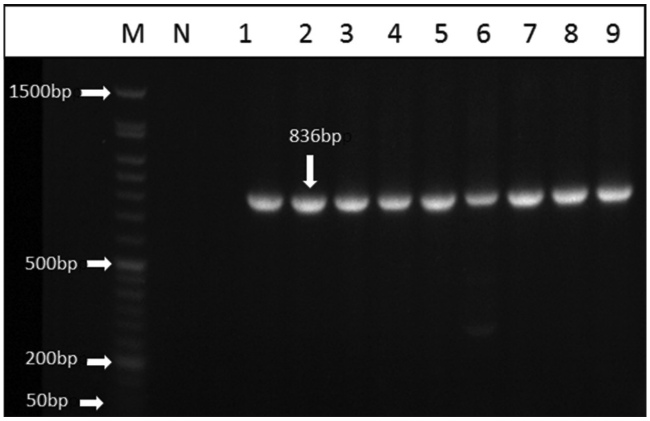


Fig. 2. Gel electrophoresis of the PCR product of the primer FCOX1. M = DNA ladder (50 bp), N = Negative control, 1–9 = The samples.

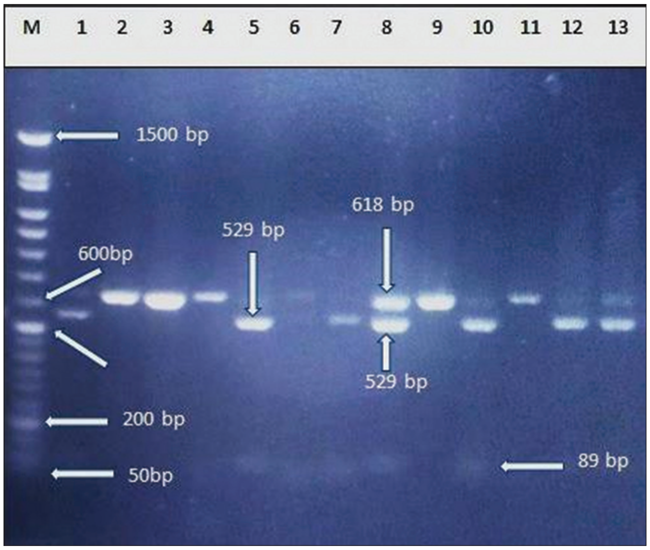


Fig. 3. Restriction fragment length polymorphism Patterns of PCR products of 28s rDNA primer digested with DraII enzyme: M = DNA Ladder (50 bp), 1–13 = The samples isolated from animals (1–11) and humans (12 and 13).

Table 2: Restriction fragment length polymorphism Patterns of PCR products of 28S rDNA primer digested with DraII enzyme

Host	No. of sample	Primer name	Restriction enzyme	PCR product	Restriction fragment (size)	Fasciola spp.
Human	3	28S rDNA	Dra II	618bp	529, 89 bp	*F. hepatica*
Sheep	20	28S rDNA	Dra II	618bp	529, 89 bp 618 bp	*F. hepatica* *F. gigantica*
Cattle	20	28S rDNA	Dra II	618bp	529, 89 bp 618 bp	*F. hepatica* *F. gigantica*
Goat	5	28S rDNA	Dra II	618bp	529, 89 bp 618 bp	*F. hepatica* *F. gigantica*

F. hepatica: Fasciola hepatica, PCR: Polymerase chain reaction

4. DISCUSSION

The present study is an attempt to evaluate two DNA markers to find the genetic characterization and species identification of liver flukes isolated from humans and animals, the RFLP and DNA sequencing were used for two DNA regions of the liver flukes. There are no adequate studies on the molecular and species identification of the liver fluke isolated from humans in Kurdistan Region – Iraq. In the present study, RFLP techniques were applied to identify the *Fasciola* spp. Many studies have been published in various countries throughout the world to identify the genetic polymorphism and species identification of this fluke, including Iraq, Iran, and Thailand [12], [13], [15]. In a study on the molecular diagnosis of *Fasciola hepatica* in livestock utilizing the COX1 gene in Erbil Province, Kurdistan – Iraq, PCR, and sequencing were used to identify the genetic variation among isolated liver fluke from animals they found genetic variation and they conclude that DNA sequencing is a differential technique for finding genetic variation and species identification using COX1 region, this finding in agreement with result [12]. In another study in Duhok governorate – Iraq, specifically used the ribosomal DNA markers; ITS1, and they found the common species that are responsible for animal fascioliasis are *F. hepatica* and *F. gigantica* this result agrees with the current finding [15].

Despite earlier studies, recent investigation on the prevalence of fascioliasis and molecular characterization of *Fasciola* spp. in sheep and goats was done in the Sulaymaniyah province of northern Iraq, using DNA markers and sequencing of the incomplete 28S rRNA gene and codon analysis, they found that *F. hepatica* represented the majority of the identified, followed by *F. gigantica* represented the other two field sequences, this study also supports the finding of the present study and they confirmed the Sulaymaniyah is the place to both *F. hepatica* and *F. gigantica*. the 28S rDNA is confirmed as a potential biomarker in identifying various *Fasciola* species, it was also used in the current study [9].

Furthermore, many studies from different areas demonstrate that molecular approaches are applicable to discriminate *Fasciola* species, the commonly used restriction enzymes Ava II and Dra II, based on a sequence of the 28S rRNA gene, *COX1, ITS1, and ITS2* [15], [16], [17], [18]. They concluded that RFLP is rapid and easy for species identification of liver flukes, and they support current results that indicated the incubation period of DNA digestion by the restriction enzymes can be shortened, also different enzymes can be used in future studies to find the most applicable enzyme to discriminate all three species of *Fasciola*.

Despite the use of genetic markers for *Fasciola* species identification, the immunological markers were done to find the efficacy and applicability to discriminate the species of liver flukes, a study from Duhok, Kurdistan – Iraq demonstrated the effectiveness of an enzyme-linked immunosorbent assay (ELISA) in infected sheep with *Fasciola* spp. the direct examination visual inspection of the liver and immunological assay was followed [7], the result shows the infection rate among animals at a high rate while the species of liver flukes cannot be accurately diagnosed immunologically due to morphologically high similarity, this make researches to adapt molecular markers exactly PCR-PFLP which is the dependable approach for the detection of *Fasciola* species.

5. CONCLUSION

The our finding concluded that all species of *Fasciola* were found in studied area and *F. hepatica* is the common species responsible of human infection. Furthermore, the present study concludes the use of molecular markers PCR-RFLP with different restriction enzymes for species identification and genetic variation is rapid, easy, and reliable. Furthermore, nucleotide sequencing can be used for species identification and genotyping while the nucleotide sequence analysis and editing will make mistakes or substation of the nucleotide base pair and the result may be not accurate as in the PCR-RFLP which is depend on the restriction side and it will be highly specific.

6. ACKNOWLEDGMENT

The authors would like to deeply thank the Sulaimani Polytechnic University, Human Development University, Veterinary Directorate of Sulaimani, and Sulaimani Health Directorate/Center for gastroenterology and hepatology in Sulaimani for their logistic support and facilities.

REFERENCES

[1] Ö. Y. Çelik and B. A. Çelik. "Investigation of the prevalence of *Fasciola hepatica* in small ruminants in the Siirt region, Turkey". *Iranian Journal of Parasitology*, vol. 13 no. 4, pp. 627-631, 2018.

[2] K. Piri, M. Saidijam, A. Maghsood, M. Matini and M. Fallah. "Prevalence of animal Fasciolosis and specification of *Fasciola* spp. isolated from sheep, goats and cattle, by molecular method: Hamadan province, West of Iran". *Iranian Journal of Parasitology*, vol. 13, no. 4, pp. 524-531, 2018.

[3] R. A. da Costa, L. G. Corbellini, E. Castro-Janer and F. Riet-Correa. "Evaluation of losses in carcasses of cattle naturally infected with *Fasciola hepatica*: Effects on weight by age range and on carcass quality parameters". *International Journal of Parasitology*, vol. 49, no. 11, pp. 867-872, 2019.

[4] J. C. Pinilla, A. A. F. Muñoz and N. U. Delgado. "Prevalence and risk factors associated with liver fluke *Fasciola hepatica* in cattle and sheep in three municipalities in the Colombian Northeastern Mountains". *Veterinary Parasitology, Regional Studies and Reports*, vol. 19, p. 100364, 2020.

[5] N. A. Ouchene-Khelifi, N. Ouchene, H. Dahmani, A. Dahmani, M. Sadi and M. Douifi. "Fasciolosis due to *Fasciola hepatica* in ruminants in abattoirs and its economic impact in two regions in Algeria". *Tropical Biomedicine*, vol. 35, no.1, pp. 181-187. 2018.

[6] Y. Zhang, H. Xu, Y. Liu, J. Kang, H. Chen, Z. Wang and D. Cai. "Case report : Fascioliasis *Hepatica* precisely diagnosed by metagenomic next-generation sequencing and treated with albendazole". *Frontiers in Medicine* (*Lausanne*), vol. 8, p. 773145, 2021.

[7] A. A. Meerkhan and A. H. Razak. "The differences between direct examination and enzyme linked immunosorbent assay (ELISA) test, during the diagnosis of fasciolosis in Jaundiced slaughtered sheep in Duhok Abattoir, Kurdistan region of Iraq". *International Journal of Chemical, Environmemtal and Biological Sciences*, vol. 1, no. 5, pp. 707-709, 2013.

[8] D. N. Anh, L. T. Anh, L. Q. Tuan, N. D. Bac, T. V. Tien, V. T. B. Phuong, T. T. Duong, N. K. Luc and N. B. Quang. "Identification of *Fasciola species* isolates from Nghe an province, Vietnam, based on ITS1 sequence of ribosomal DNA using a simple PCR-RFLP method". *Journal of Parasitology and Research*, vol. 20, no. 2018, p. 2958026, 2018.

[9] H. S. Rahman, H. Marif, M. O. B. Sheikh and A. M. Ahmed. "Molecular characterization and phylogenetic analysis of *Fasciola* species in sheep and goats in Sulaymaniyah province, Northern Iraq". *Journal of Zankoy Sulaimani Part A*, vol. 22, pp. 297-305, 2020.

[10] D. Teofanova, V. Kantzoura, S. Walker, G. Radoslavov, P. Hristov, G. Theodoropoulos, I. Bankov and A. Trudgett. "Genetic diversity of liver flukes (*Fasciola hepatica*) from Eastern Europe". *Infection Genetics and Evolution*, vol. 11, no. 1, pp. 109-115, 2011.

[11] A. E. Laatamna, M. Tashiro, Z. Zokbi, Y. Chibout, S. Megrane, F. Mebarka and M. Ichikawa-Seki. "Molecular characterization and phylogenetic analysis of *Fasciola hepatica* from high-plateau and steppe areas in Algeria". *Parasitology International*, vol. 80, 2021, p. 102234.

[12] M. J. Muhammad and Z. I. Hassan. "Molecular diagnosis of *Fasciola hepatica* in livestock using cox1 gene in Erbil province-Kurdistan region/Iraq". *Zanco Journal of Pure and Applied Sciences*, vol. 33, no. 4, pp. 36-42. 2021.

[13] M. Aryaeipour, S. Rouhani, M. Bandehpour, H. Mirahmadi, B. Kazemi and M. B. Rokni. "Genotyping and phylogenetic analysis of *Fasciola* spp. isolated from sheep and cattle using PCR-RFLP in Ardabil province, Northwestern Iran". *Iranian Journal of Public Health*, vol. 43, No. 10, pp. 1364-1371, 2014.

[14] A. Marcilla, M. D. Bargues and S. Mas-Coma. "A PCR-RFLP assay for the distinction between *Fasciola hepatica* and *Fasciola gigantica*". *Molecular and Cellular Probes*, vol. 16, no. 5, pp. 327-333, 2002.

[15] P. Siribat, P. Dekumyoy, C. Komalamisra, S. Sumruayphol and U. Thaenkham. "Molecular identification of *Fasciola* spp. -representative samples from Thailand based on PCR-RFLP". *Journal of Tropical Medicine and Parasitology*, vol. 41, no. 1, pp. 1-7, 2018.

[16] S. Farjallah, B. Ben Slimane, C. M. Piras, N. Amor, G. Garippa and P. Merella. "Molecular characterization of *Fasciola hepatica* from Sardinia based on sequence analysis of genomic and mitochondrial gene markers". *Experimental Parasitology*, vol. 135, no. 3, pp. 471-478, 2013.

[17] S. Kasahara Y. Ohari, S. Jin, M. Calvopina, H. Takagi, H. Sugiyama and T. Itagaki. "Molecular characterization revealed *Fasciola* specimens in Ecuador are all *Fasciola hepatica*, none at all of *Fasciola gigantica* or parthenogenic *Fasciola* species". *Parasitology Inernational*, vol. 80, p. 102215, 2021.

[18] A. Saadatnia, K. Solhjoo, M. H. Davami, S. Raeghi and A. Abolghazi. "Molecular identification of *Fasciola* isolated from the liver of meat animals in Fars province, Iran". *Journal of Parasitology Research*, vol. 2022, p. 4291230, 2022.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Limitations of Load Balancing and Performance Analysis Processes and Algorithms in Cloud Computing

Asan Baker Kanbar[1,2]*, Kamaran Faraj[3,4]

[1]Technical College of Informatics, Sulaimani Polytechnic University , Sulaimani 46001, Kurdistan Region, Iraq,
[2]Department of Computer Science,Cihan University Sulaimaniya, Sulaimaniya 46001, Kurdistan Region, Iraq,
[3]Department of Computer Science, University of Sulaimani, Sulaimani, 46001, Kurdistan Region, Iraq, [4]Department of Computer Engineering, Collage of Engineering and Computer Science, Lebanse Frence University, Erbil, Iraq

ABSTRACT

In the modern IT industry, cloud computing is a cutting-edge technology. Since it faces various challenges, the most significant problem of cloud computing is load balancing, which degrades the performance of the computing resources. In earlier research studies, the management of the workload to address all resource allocation challenges that caused by the participation of a large number of users has received important attention. When several people are attempting to access a given web application at once, managing all of those users becomes exceedingly difficult. One of the elements affecting the performance stability of cloud computing is load balancing. This article evaluates and discusses load balancing, the drawbacks of the numerous methods that have been suggested to distribute load among nodes, and the variables that are taken into account when determining the best load balancing algorithm.

Index Terms: Cloud Computing, Load Balancing, Task Scheduling, Resource Allocation, Task Allocation, Performance Stability

1. INTRODUCTION

Cloud computing is a new technology for a large-scale environments. Hence, it faces many challenges and the main problem of cloud computing is load balancing which lowering the performance of the computing resources [1]. Management is the key to balancing performance and management costs along with service availability. When Cloud Data Centres (CDCs) are configured and utilized effectively, they offer huge benefits of computational power while reducing cost and saving energy. Cloud computing has three types of services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Fundamental resources can be accessed through IaaS. PaaS provides the application runtime environment, besides development and deployment tools. SaaS enables the provision of software applications as a service to end users. Virtual entities are created for all hardware infrastructure elements. Virtualization is a technique that allows multiple operating systems (OSs) to coexist on a single physical machine (PM). These OSs are separated from one another and from the underlying physical infrastructure by a special middleware abstraction known as virtual machine (VM). The software that manages these multiple VMs on PM is known as the VM kernel [2]. With the help of virtualization technology, CDCs are able to share a few HPC resources and their services among many users, but virtualization

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp71-77	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Asan Baker Kanbar, Assistant Lecturer, Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya 46001, Kurdistan Region, Iraq, Asan. E-mail: asan.baker@sulicihan.edu.krd

Received: 21-11-2022 Accepted: 02-03-2023 Published: 18-03-2023

increases the complexity of resource management. Task scheduling is one of the key issues considered for efficient resource management. It aims to allocate incoming task(s) to available computing resources, and it belongs to the set of NP-class problems. Therefore, heuristic and meta-heuristic-based approaches are commonly used to generate scheduling solutions while optimizing one or more goals such as makespan, resource utilization, number of active servers, through-put, temperature effects, energy consumption, etc. Customers in the cloud can access resources at any time through the web and only pay for the services they use. With the dramatic increase in cloud users, decreasing task completion time is beneficial for improving user experience. The primary goals of task scheduling are to reduce task completion time and energy consumption while also improving resource utilization and load balancing ability [3]. Improving load balancing ability contributes to fully utilizing VMs to prevent execution efficiency from decreasing due to resource overload or waste caused by excessive idle resources. Various algorithms have been proposed to balance the load between multiple cloud resources, but there is currently no algorithm that can balance the load in the cloud without degrading performance. Load balancing is a method used to improve the performance of networking by distributing the workload among the various resources involved in computing network tasks. The load here can be processor capacity, memory, network load, etc.

Load balancing optimizes resource usage, reduces response time, and avoids system overload by distributing the load across several components. Many researchers are working on the problem of load balancing, and as a result of their research, many algorithms are proposed every day. In this paper, we overview some of the optimistic algorithms that have shown some improvement in load balancing and increased the level of performance. Besides we will also show the limitations of these algorithms.

2. LOAD BLANCING IN CLOUD COMPUTING

Load balancing is performed for resource allocation and managing load in each data center, as illustrated in Fig. 1. Load balancing in a cloud computing environment has a significant impact on performance; good load balancing can make cloud computing more efficient and improve user satisfaction. Load balancing is a relatively new technology that allows networks and resources to deliver maximum throughput with a minimum response time.

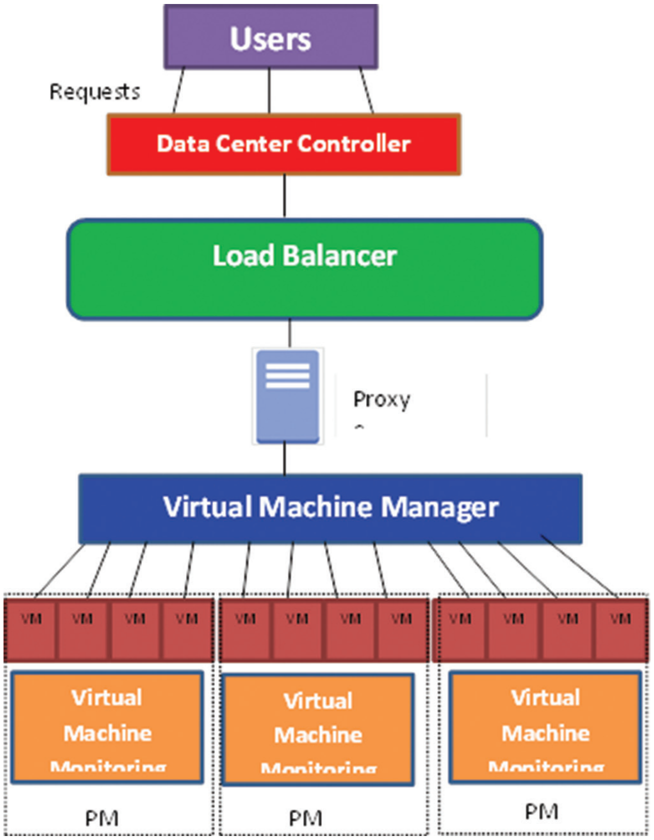


Fig. 1. Model of load balancing.

Good load balancing helps to optimize the use of available resources, thus minimizing resource consumption. By sharing traffic between servers, you can send and receive data without experiencing significant delays. Different types of algorithms can be used to help reduce traffic load between available servers. A basic example of load balancing in everyday life can be related to websites. Without load balancing, users may experience delays, timeouts, and the system may become less responsive. By dividing the traffic among servers, data can be sent and received without significant delay.

Load balancing is done using a load balancer (Fig. 2), where each incoming request is redirected and transparent to the requesting client.

Based on specified parameters such as availability and current load, the load balancer uses various scheduling algorithms to find which server should handle the request and sends the request to the selected server. To make a final decision, the load balancer obtains information about the candidate server's state and current workload to validate its ability to respond to this request [4].

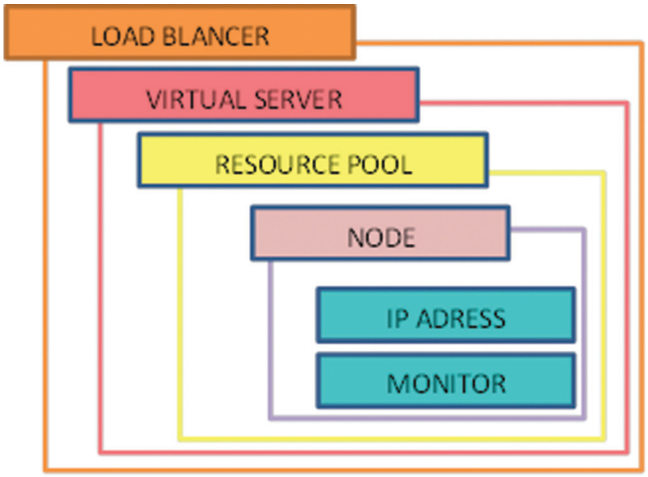


Fig. 2. Load Balancer.

3. CHALLENGES IN CLOUD COMPUTING LOAD BALANCING

Before we could review the current load balancing approaches for cloud computing, we need to identify the main issues and challenges involved and that could affect how the algorithm would perform. Here, we discuss the challenges to be addressed when attempting to propose an optimal solution to the issue of load balancing in Cloud Computing. These challenges are summarized in the following points.

3.1. Cloud Node Distribution

Many algorithms have been proposed for load balancing in cloud computing; among them, some algorithms can provide efficient results in small networks or networks with nodes close to each other. Such algorithms are not suitable for large networks because they cannot produce the same efficient results when applied to larger networks. The development of a system to regulate load balancing while being able to tolerating significant delays across all the geographical distributed nodes is necessary [5]. However, it is difficult to design a load balancing algorithm suitable for spatially distributed nodes. Some load-balancing techniques are designed for a smaller area where they do not consider the factors such as network delay, communication delay, distance between the distributed computing nodes, distance between user and resources, and so on. Nodes located at very distant locations are a challenge, as these algorithms are not suitable for this environment. Thus, designing load-balancing algorithms for distantly located nodes should be taken into account [6]. It is used in large-scale applications such as Twitter and Facebook. The DS of the processors in the cloud computing environment is very useful for maintaining system efficiency and handling fault tolerance well. The geographical distribution has a significant impact on the overall performance of any real-time cloud environment.

3.2. Storage/Replication

A full replication algorithm does not take efficient storage utilization into account. This is because all replication nodes store the same data. Full replication algorithms impose higher costs because more storage capacity is required. However, partial replication algorithms may store partial datasets in each node (with some degree of overlap) depending on each node's capabilities (such as power and processing capacity) [7]. This can lead to better usability, but it increases the complexity of load balancing algorithms as they try to account for the availability of parts of the dataset on different cloud nodes.

3.3. Migration Time

Cloud computing follows a service-on-demand model, which means when there is a demand for a resource, the service will be provided to the required client. Therefore, while providing services based on the needs of our customers, we sometimes have to migrate resources from remote locations due to the unavailability of nearby locations. In such cases, the time of migration of the resources from far locations will be more which will affect system performance. When developing algorithms, it is important to note that resource migration time is an important factor affecting system performance.

3.4. Point of Failure

Controlling the load balancing and collecting data about the various nodes must be designed in a way that avoids having a single point of failure in the algorithm. If the algorithm's patterns are properly created, they can also help provide effective and efficient techniques to address load balancing problems. Using a single controller to balance the load is a major difficulty because, failure might have severe consequences and lead to overloading and under-loading issues. This difficulty must be addressed in the design of any load balancing algorithm [8]. Distributed load balancing algorithms seem to offer a better approach, but they are much more complex and require more coordination and control to work properly.

3.5. System Performance

This does not mean that if the complexity of the algorithm is high, then the system performance will be very high. Any time a load balancing algorithm must be simple to implement and easy to operate. If the complexity of the algorithm is high, then the implementation cost will also be higher and even after implementing the system, performance will be decreased due to the increased delays in the functionality of the algorithm.

3.6. Algorithm Complexity

In terms of implementation and operation, the load balancing algorithm is preferably not that complicated. Higher implementation complexity will lead to more complex procedures, which can lead to negative performance issues Furthermore, when the algorithms require more information and higher communication for monitoring and control, delays would cause more problems and reduce efficiency. Therefore, to reduce overhead on cloud computing services, load-balancing algorithms should be as simple and effective as possible [9].

3.7. Energy Management

A load balancing algorithm should be designed in such a way that the operational cost and energy consumption of the algorithm must be low. Increasing energy consumption is one of the biggest issues facing cloud computing today. Even though using energy efficient hardware architectures which slows down the processor speed and turn off machines that are not under use the energy management is becoming difficult. Hence, to achieve better results in energy management, the load balancing algorithm should be designed according to Energy Aware Job Scheduling methodology [10].

3.8. Security

Security is one of the problems that cloud computing has as its top priority. The cloud is always vulnerable in one way or the other way to security attacks like DDOS attacks, etc. While balancing the load there are many operations that take place like VM migration, etc. at that time there is a high probability of security attacks. Hence, an efficient load balancing algorithm must be strong enough to reduce security attacks but should not be vulnerable.

4. RELATED WORKS AND LIMITATIONS OF USED ALGORITHMS AND PROCESSES

The author [11] proposed a hybrid optimization algorithm for load balancing. This is firefly optimization and enhanced multi-criteria based on the Particle Swarm Optimization (PSO) algorithm called (FIMPSO). To initialize the population in PSO, the Firefly algorithm is used, since it gives the optimal solution. Only two parameters are considered here, such as task arrival time and task execution time. The results are executed, taking into account parameters such as run time, resource consumption, reliability, makespan, and throughput. Limitations: Hybrid algorithms require high latency to run. In particular, PSO falls into a local optimum problem when processing a large number of requests, and the convergence speed is low. Overloading occurs here because more iteration is needed to achieve the optimal solution. In the paper [12], propose the use of three-layer cooperative fog to reduce bandwidth cost and delay in cloud computing environments, this article discusses the composite objective function of bandwidth cost reduction and load balancing, where we consider both link bandwidth and server CPU processing levels. Assign weights to every objective of the composite objective function to determine priority. The minimum bandwidth cost has a higher priority and runs first on Layer1 fog. However, the load balancer gets the priority it used to reduce latency. The MILP (Mixed-Integer Linear Programming) algorithm is used to minimize the composite objective function. Two types of resources are used, one is a network resource (bandwidth) and the other is a server resource (CPU processing layer). Limitations: This work is not suitable for real-time applications, because it takes a high execution time for selecting the bandwidth and CPU. It only focuses on reducing bandwidth costs and load balancing, so it takes a long time to find the optimal solution. Priority is based on the minimum bandwidth utilization in a large scale environments, many regions are used the minimum bandwidth utilization so congestion is occurring; it takes much time to execute the task, which also reduces the QoS values. Author [13] Task offloading and resource allocation were proposed for IoT fog cloud architecture based on energy and time efficiency. The ETCORA algorithm is used to improve energy efficiency and request completion time. It performs two tasks. One is computational offload selection and the other is transmitting power allocation. Three layers are presented in this work. The first tier contains some IoT devices. The second tier is the fog tier, which consists of fog servers and controllers located in different geographic locations. The third tier is the cloud tier, which consists of cloud servers. However, the entire task is outsourced within the fog layer, so the fog layer is also overloaded. In many regions, a request is sent to the users at a certain time, the fog layer cannot control the load balancing. All users in the region access the cloud server, which triggers load balancing. The author [14] proposed using probabilistic load balancing to avoid congestion due to VM migration and also to minimize congestion across migrations. For VM migration, this paper takes into account the distance between the source PM and the destination PM. The architecture features a VM migration controller, stochastic demand forecasting, hotspot detection, and VMs, PMs. Load balancing is addressed by profiling resource demand, hotspot demand, and hotspot migration. Resource demand profiling tracked the following: VM resource utilization on CPU, memory, network bandwidth, and disk I/O. It is used to update the periodic information to the balancer. For discovering the hotspot they periodically change the resource allocation status from the VMs and PMs' Resource demands. The hotspot migration process uses the hotspot migration algorithm.

Author [15] proposed a static load balancing algorithm totally based on discrete PSO for distributed simulations in cloud computing. For static load balancing, adaptive pbest discrete PSO (APDPSO) is used. PSO updates particle velocity and position vectors. The distance metric is used to update the velocity and position vectors from the pbest and gbest values. Non-Dominated Genetic Sorting Algorithm II (NSGA II) is one of the evolutionary algorithms that preserves the optimal solution. For each iteration, NSGA II considers three important processes selection, mutation and crossover. However, PSO suffers from local optima and poor convergence when handling a large number of requests, resulting in increased latency. In paper [16], the author proposed multi-goal task scheduling based on SLA and processing time, which is suitable for cloud environments. This article proposes two scheduling algorithms called the Threshold Based Task Scheduling (TBTS) algorithm and the Service Level Agreement Load Balancing (SLA-LB) algorithm. TBTS scheduled a task for a batch TNTS threshold (expected time of completion) generated from ETC. SLA-LB is based on an online model that dynamically schedules a task based deadline and budget criteria. SLA-LB is used to find the required system to reduce the makespan and increasing the cloud usage. This paper discuses following performance metrics such as makespan, penalty, achieve cost, and VM utilization. The results are shows that the proposed method is superior when compared to existing algorithm in terms of both scalability and VMs. However, the value of threshold is based on the completion time, if assuming that completion time is increased, the threshold value will be burst. It reduces the SLA and QoS values. Author [17] proposed a multi-agent system for dynamic consolidation of VMs with optimized energy efficiency in cloud computing. This proposed system eliminates the centralized failure, so that, the decentralized server presented with Gossip Control (GC) with a multi-agent framework the GC has two protocols: Gossip and Contract Network Protocol. With the assist of GC developed DVMS (Dynamic VM Consolidation) compared two sercon strategies centralized strategy and an ECO Cloud distributed Strategy. Sercon is used to minimize server count and VM migration. During integration, Eco Cloud considers two processes: First is the migration procedure and the second is the allocation procedure GC-based strategy works best for SLA violations and power consumption. In paper [18] using cloud theory for wind power uncertainty, the author proposed a multi-objective feeder reconfiguration problem. Proposed used the cloud theory properties of qualitative–quantitative bidirectional transmission to solve the problem of multi-objective feeder reconfiguration with the backward and forward cloud generator algorithm, Proposed system used a fuzzy decision-making algorithm to get the best solution. Authors in [19] proposed an approach to perform cloud resource provisioning and scheduling based on metaheuristic algorithms. To design the supporting model for the autonomic resource that schedules the applications effectively, the binary PSO (BPSO) algorithm is used. This work consists of three consecutive phases as user-level phase, the cloud provisioning and scheduling phase, and the infrastructure level phase. Finally, experimental evaluation was performed by modifying the BPSO algorithm's transfer function to achieve high exploration and exploitation. Authors in [20] introduced intelligent scheduling and provisioning of resource methods for cloud computing environments. This work overcomes the existing problems of poor quality of service, increased execution time, and high costs during service. Existing problems are addressed by an intelligent optimization framework that schedules jobs for users using spider monkey optimization. This will result in faster execution time, lower cost, and better QoS for the user. Here, the job is scheduled by the Spider-Monkey optimization algorithm. However, Job sensitivity (i.e., risk or non-risk) is not considered, resulting in poor QoE. Author [21] proposed to consider a multi-objective optimization for energy-efficient allocation of virtual clusters in CDCs. This research paper describes four optimization goals related to VC and data centers: Availability, power consumption, average resource utilization, and resource load balancing. The architecture contains three-layers which are the core layer, aggregation layer, and edge layer. In the core, layer contains a core switch, which is classified into many aggregated switches. The aggregation layer contains an edge switch and PMs. The edge switches are connected to the PMs. The PMs are connected to the VMs cluster. If the edge switch is failed, it will not access the PMs and VMs cluster.

5. DISCUSSION

The existing works were addressed the issues of task scheduling and load balancing in IoT-fog-cloud environment. Many of the

TABLE 1: Summary of Related Works and Limitations of Used Algorithms and processes

References	Task classification	Task scheduling	Load balancing	Task allocation	Algorithm/process used	limitations
[11]	x	x	✓	x	Firefly improved multi-objective particle swarm optimization (FIMPSO).	• More number of iterations and low convergence rate Performs
[12]	x	x	✓	x	When traffic exceeds a region's capacity, the fog layer performs load balancing.	• Long execution time is observed • Task requirements were not considered resulting in SLA violation.
[13]	x	x	✓	✓	ETCORA algorithm for energy efficient offloading	• Constraints exist when it comes to increasing the scalability of tasks in a particular region.
[14]	x	x	✓	x	Resource requirements for each task are profiled and offloaded.	• Throughput is affected because of frequent migrations • High migration time because of long congestion in link.
[15]	x	x	✓	x	Load balancing based on Adaptive Pbest Discrete PSO (APDPSO) to reduce communication costs	• Blind spot problem • Throughput is affected because of frequent migrations
[16]	x	✓	✓	x	TBTS and SLA-LB to dynamically perform scheduling and load balancing.	• Ineffective determination of threshold value increases complexity
[17]	x	x	✓	x	Gossip Control based Dynamic Virtual machine Consolidation for effective load balancing	• Lack of security of data during migration. • High migration time because of long congestion in link.
[18]	x	✓	x	✓	BPSO based resource provisioning and scheduling	• High privacy leakage during data migration
[19]	x	x	✓	x	Cloud theory based optimization of load in order improve QoS	• Increased latency and end to end delay because centralized processing in cloud layer
[20]	x	✓	x	✓	ARPS framework based scheduling and provisioning of resources	• High congestion occurs due data overloading
[21]	x	x	x	✓	Virtual clustering based multi objective task allocation is carried out using IBBBO.	• long execution time • Slow convergence of IBBBO algorithm

research aims to reduced makespan, energy consumption, and latency during task scheduling, allocation, and VM migration for load balancing. However, the existing works consider limited features for tasks scheduling and allocation which leads to poor scheduling and QoS. In addition, some of the works selects target VM for migration by considering only load which was not enough for optimal VM migration. Due to lack of significant features also increases frequent migration, which increase high overload and latency during load balancing. The existing works used optimization algorithm with slow convergence such as PSO, genetic algorithm, etc. for task scheduling and VM migration which leads to high latency and overload during load balancing in IoT-fog-cloud environment; hence, we need to addressed these issues for providing efficient task scheduling and load balancing results.

6. CONCLUSION

Cloud computing is growing rapidly and users are demanding more and more services, that's why cloud computing load balancing has become such a thoughtful impetus and important research area. Load on the cloud is growing extremely with the expansion of new applications [22] to overcome the load because of huge requests and increase the quality of service many load balancing techniques are used. In this paper, we surveyed many cloud load balancing techniques and focusing to the limitations of each to help the researchers to propose new methods to overcome the limitations to solve the problem of load balancing in cloud computing environment. Table 1 shows the summary of related works the used load balancing algorithms and processes limitations.

REFERENCES

[1] A. B. Kanbar and K Faraj. “Regional aware dynamic task scheduling and resource virtualization for load balancing in IoT-Fog multi-cloud environment”. *Future Generation Computer Systems*, 137C, pp. 70-86, 2022.

[2] H. Nashaat, N. Ashry and R. Rizk. “Smart elastic scheduling algorithm for virtual machine migration in cloud computing”. The Journal of Supercomputing, vol. 75, pp. 3842-3865, 2019.

[3] H. Gao, H. Miao, L. Liu, J. Kai and K. Zhao. “Automated quantitative verification for service-based system design: A visualization transform tool perspective”. *International Journal of Software Engineering and Knowledge Engineering*, vol. 28, no. 10, pp. 1369-1397, 2018.

[4] S. V. Pius and T. S. Shilpa. “Survey on load balancing in cloud computing”. In: *International Conference on Computing, Communication and Energy Systems*, 2014.

[5] P. Jain and S. Choudhary. “A review of load balancing and its challenges in cloud computing”. *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 4, pp. 9275-9281, 2017.

[6] P. Kumar and R. Kumar. “Issues and challenges of load balancing techniques in cloud Computing: A survey”. *ACM Computing Surveys*, vol. 51, no. 6, pp. 1-35, 2019.

[7] M. Alam and Z. A. Khan. “Issues and challenges of load balancing algorithm in cloud computing environment”. *Indian Journal of Science and Technology*, vol. 10, no. 25, pp. 1-12, 2017.

[8] H. Kaur and K. Kaur. “*Load Balancing and its Challenges in Cloud Computing: A Review*”. In: M. S. Kaiser, J. Xie and V. S. Rathore, editors. Information and Communication Technology for Competitive Strategies (ICTCS 2020). Lecture Notes in Networks and Systems, vol. 190. Springer, Singapore, 2021.

[9] G. K. Sriram. “Challenges of cloud compute load balancing algorithms”. *International Research Journal of Modernization in Engineering Technology and Science*, vol. 4, no. 1, p. 6, 2022.

[10] H. Chen, F. Wang, N. Helian and G. Akanmu. “*User-priority Guided Min-min Scheduling Algorithm for Load Balancing in Cloud Computing*”. In: Proceeding National Conference on Parallel Computing Technologies (*PARCOMPTECH*), IEEE. pp. 1-8, 2013.

[11] A. F. Devaraj, M. Elhoseny, S. Dhanasekaran, E. L Lydia and K. Shankar. “Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments”. *Journal of Parallel and Distributed Computing*, vol. 142, pp. 36-45, 2020.

[12] M. M. Maswood, M. R. Rahman, A. G. Alharbi and D. Medhi. “A novel strategy to achieve bandwidth cost reduction and load balancing in a cooperative three-layer fog-cloud computing environment”. *IEEE Access*, vol. 8, pp. 113737-113750, 2020.

[13] H. Sun, H. Yu, G. Fan and L. Chen. “Energy and time efficient task offloading and resource allocation on the generic IoT-fog-cloud architecture”. *Peer-to-Peer Networking and Applications*, vol. 13, no. 2, pp. 548-563, 2020.

[14] L. Yu, L. Chen, Z. Cai, H. Shen, Y. Liang and Y. Pan. “Stochastic load balancing for virtual resource management in datacenters”. *IEEE Transactions on Cloud Computing*, vol. 8, pp. 459-472, 2020.

[15] Z. Miao, P. Yong, Y. Mei, Y. Quanjun and X. Xu. “A discrete PSO-based static load balancing algorithm for distributed simulations in a cloud environment”. *Future Generation Computer Systems*, vol. 115, no. 3, pp. 497-516, 2021.

[16] D. Singh, P. S. Saikrishna, R. Pasumarthy and D. Krishnamurthy. “Decentralized LPV-MPC controller with heuristic load balancing for a private cloud hosted application”. *Control Engineering Practice*, vol. 100, no. 4, p. 104438, 2020.

[17] N. M. Donnell, E. Howley and J. Duggan. “Dynamic virtual machine consolidation using a multi-agent system to optimise energy efficiency in cloud computing”. *Future Generation Computer Systems*, vol. 108, pp. 288-301, 2020.

[18] F. Hosseini, A. Safari and M. Farrokhifar. “Cloud theory-based multi-objective feeder reconfiguration problem considering wind power uncertainty”. *Renewable Energy*, vol. 161, pp. 1130-1139, 2020.

[19] M. Kumar, S. C. Sharma, S. S. Goel, S. K. Mishra and A. Husain. “Autonomic cloud resource provisioning and scheduling using meta-heuristic algorithm”. *Neural Computing and Applications*, vol. 32, pp. 18285-18303, 2020.

[20] M. Kumar, A. Kishor, J. Abawajy, P. Agarwal, A. Singh and A. Y. Zomaya. “ARPS: An autonomic resource provisioning and scheduling framework for cloud platforms”. *IEEE Transactions on Sustainable Computing*, vol. 7, no. 2, pp. 386-399, 2021.

[21] X. Liu, B. Cheng and S. Wang. “Availability-aware and energy-efficient virtual cluster allocation based on multi-objective optimization in cloud datacenters”. *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 972-985, 2020.

[22] A. B. Kanbar and K. Faraj. “Modern load balancing techniques and their effects on cloud computing”. *Journal of Hunan University* (*Natural Sciences*), vol. 49, no.7, pp. 37-43, 2022.

SURVEY UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Missing Value Imputation Techniques: A Survey

Wafaa Mustafa Hameed[1,2]*, Nzar A. Ali[2,3]

[1]Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, 46001, Kurdistan Region, Iraq,
[2]Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya, 46001, Kurdistan Region, Iraq,
[3]Department of Statistics and informatics, University of Sulaimani, Sulaimani, 46001, Kurdistan Region, Iraq

ABSTRACT

Numerous of information is being accumulated and placed away every day. Big quantity of misplaced areas in a dataset might be a large problem confronted through analysts due to the fact it could cause numerous issues in quantitative investigates. To handle such misplaced values, numerous methods were proposed. This paper offers a review on different techniques available for imputation of unknown information, such as median imputation, hot (cold) deck imputation, regression imputation, expectation maximization, help vector device imputation, multivariate imputation using chained equation, SICE method, reinforcement programming, non-parametric iterative imputation algorithms, and multilayer perceptrons. This paper also explores a few satisfactory choices of methods to estimate missing values to be used by different researchers on this discipline of study. Furthermore, it aims to assist them to discern out what approach is commonly used now, the overview may additionally provide a view of every technique alongside its blessings and limitations to take into consideration of future studies on this area of study. It can be taking into account as baseline to solutions the question which techniques were used and that is the maximum popular.

Index Terms: Data Preprocessing, Imputation, Mean, Mode, Categorical Data, Numerical Data

1. INTRODUCTION

Data mining has made an amazing development in the past years; however, the main problem is missing data or value. Data mining is the sector wherein experimental facts sets are analyzed to find out thrilling and potentially beneficial relationships [1]. Lacking records or value in a datasets can affect the performance of classifier which ends up in difficulty of extracting beneficial information from datasets. Plentiful of facts is being gathered and saved each day. Those facts can be used to extract interesting patterns. The information that we collect is incomplete normally [2]. Therefor everyone wishing to apply statistical information evaluation or information cleaning of any type could have problems with lacking data. We still land in some missing attribute values in a function dataset. People typically tend to depart the income area empty in surveys, for instance, and members once in a while do not have any information available or cannot answer the question. Plenty facts can also be lost in the technique of collecting information from multiple resources [1]. Using that information to collect some statistics can now yield misleading effects. Hence, to eliminate the abnormalities, we need to pre-method the statistics earlier than the usage of it. Those instances may be omitted within the case of a small percentage of lacking values, but within the case of huge quantities, ignoring them will now not yield the desired outcome. A number of missing spaces in a dataset is a massive problem. Therefore, a few pre-processing of statistics can be accomplished earlier than acting any information mining techniques to extract a few treasured records from a dataset to keep away from such mistakes and as a result enhance statistics first-class. Fittingly managing with misplaced values is crucial and difficult venture since it requires careful examination of all occurrences of information to recognize design of missingness within the

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp72-81
E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Technical College of Informatics, Sulaimani Polytechnic University, Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya, 46001, Kurdistan Region, Iraq. E-mail id: wafaa.mustafa@spu.edu.iq

Received: 09-11-2022 Accepted: 05-03-2023 Published: 28-03-2023

data. Numerous strategies were proposed to address such lacking values considering 1980 [2].

This file illustrates distinct varieties of lacking values and the techniques used to address them.

It is tremendously vital to note that there may be evaluation in purge and lost value. Purge value implies that no value may be doled out though misplaced value implies actual value for that variable exists but not reachable or captured in dataset due to some motives. The information mineworker should separate between purge esteem and lost esteem. Once in a while, each the values may be treated as misplaced values. Lost records may be due to tools glitch, conflicting with different facts so erased, data no longer entered because of false impression, positive facts might not be considered crucial at the time of statistics collection. a few statistics mining calculations do not require substitution of misplaced values as they are planned and created to handle lost values; however, some data mining calculations cannot good buy with lost values. Sometime, these days making use of any strategy of managing with lost values its miles vital to get it why records is misplaced [2], [3].

2. MISSING VALUE PATTERNS

2.1. Missing Completely at Random (MCAR)

MCAR is the most improved degree of randomness and it indicates that the layout of misplaced value is completely arbitrary and does not rely on any variable which may additionally or might not be covered inside the examination [3]. It refers to facts that do not rely on the interest variable or every other parameter observed inside the dataset [4]. While missing values are distributed uniformly across all measurements, then we find the records to be absolutely randomly missing. For this reason, a brief test is to compare pieces of data – one with missing observations and the other without missing observations. On a t-test, if there is no mean difference between the two data units, we will expect that the data are MCAR [5]. Anything that is missing and sometimes because this form of missing facts is not often observed and the best manner to ignore these instances, for example: Water damage to paper forms due to flooding before it enters [1], [2] or in a survey, if we get 5% responses missing randomly, it is MCAR [6], [7]. This type is described by using the equation

$$P\,(\,p_1\;|\,X\,,Y_{0\,,l}\,,Y_{m\;\,,l\,)}\;=f\,(\,l\,,X\,)$$

Where f is a function, that is, the missing data patterns are determined only by the covariate variables X. Note here that MARX is equivalent to MCAR if there are no covariates in the model [7], [8].

2.2. Missing at Random (MAR)

When missed value does not rely on any given or ignored value [8]. Often information may not be deliberately missing; however, it can be named "missing at random". If the data meet the requirement that missingness should not rely on X's value after accounting for some other parameter, we may also find an X entry to be missing at random. Depressed people seem to have less income, as an instance, and the reported earnings now depend on the thing depression. The percentage of lacking records among depressed people could be high as depressed people have lower incomes [1] if we get 10% missing for the male responses in a survey and 5% missing for the woman survey, then it is MAR [6]. this kind is defined through the equation

$$P\left(p_l|X,Y_{0,l},Y_{m,l}\right)=f(l,X,Y_{0,l})$$

In which f is a function, that is, only the covariate variables X and the based variables has been located have an impact on the patterns of lacking statistics. Remember the fact that if there may be most effective one dependent variable Y then there may be best one missing series that does not encompass any found dependent variables. For models with one structured variable, MAR is therefore equal to MARX [7].

2.3. Not Missing at Random (NMAR)

If the data are not missing at random or informatively, it is labeled "not missing at random." This kind of situation happens while the technique of messiness depends at the actual value of missing statistics [4]. This type is defined by the equation:

$$P\left(p_l|X,y_{0,l},Y_{m,l}\right)=f(l,X,Y_{0,l},Y_{m,l})$$

Where f is a function, that is, all three types of variables have an effect on the missing data patterns. It is well known how full information maximum likelihood (FIML) estimation performs under all of these conditions [7].

2.4. Missing in Cluster (MIC)

Data are regularly more missing in some attributes than in others. In addition, the missing values in the ones attributes can be correlated. It is extremely tough to use statistical techniques to show multi-attribute correlations of lacking values. On this sample of missing values, the exceptional of statistics is much less homogeneous than that with MAR.

The effects of any applications of analytical based on the complete facts set have to be cautious, for the reason that pattern data are biased in the attributes with a big number of missing values [7], [8].

2.5. Systematic Irregular Missing (SIM)

Data can be missing quite irregularly, however systematically. There is probably overly missing correlations among the attributes, but those correlations are extraordinarily tiresome to analyze. An implication of SIM is that the data with complete entities are unpredictably under-representative [7]. The first-class of records with this sample of missing values is minimal homogeneous than the ones in MAR and additionally less controllable than that with MIC. Applications of any analytical results based at the whole data set are enormously questionable [9].

3. STRATEGIES OF HANDLING MISSING DATA

Managing missing data may be carried out in two exclusive strategies for. The first method is definitely ignoring missing values and second approach is to take into account imputation of missing values.

3.1. Ignoring Missing Values

The missing records ignoring technique absolutely releases the state that includes missing data. They are mightily used for handling lacking facts. The earnest problem with this method is that it decreases the dataset size. This is handy whilst the dataset has small amount of lacking values. There are two common methods for ignoring missing data:

3.1.1. Listwise deletion

Complete case analysis approach excludes all observations with missing values for any variable of interest. This approach thus limits the analysis to those observations for which all values are observed. This techniques is simple to use but cause loss of huge data, loss of precision, high effect on variability, and induce bias.

3.1.2. Pairwise deletion

For all the instances, we perform analysis with in which the variables of interest are present. It does no longer exclude complete unit but uses as lots data as feasible from every unit. This method is straightforward, keeping all available values, that is, best missing values are deleted but motive the loss of data, no longer a higher solution compared to other techniques. The pattern size for every individual evaluation is better than the entire case analysis [2], [10].

3.2. Single Imputation

Single imputation procedures produce a precise value for a dataset's missing real value. This method necessitates a lower computing cost. Researchers have proposed a variety of single imputation strategies. The typical strategy is to analyze other responses and select the greatest possible response. The value can be calculated using the mean, median, or mode of the variable's available values. Single imputation can also be done using other methods, such as machine learning-based techniques. Imputed values are considered actual values in single imputation. Single imputation ignores the reality that no imputation method can guarantee the true value. Single imputation approaches ignore the imputed values' uncertainty. Instead, in future analysis, they recognize the imputed values as actual values [11], [12].

3.3. Multiple Imputations

The use of distinct simulation models, multiple imputation methods yield several values for the imputation of single missing records. Those strategies use imputed data's variability to generate a diffusion of credible responses. Multiple imputation strategies are sophisticated in nature, but in contrast to single imputation, they do no longer suffer from bias values. In multiple imputations, every missing facts point is replaced with m values obtained through m iterations (wherein m > 1 and m generally sits between 3 and 10) [6]. In this technique, a statistical approach used for coping with missing values, it performs through three stages:

- Imputation: Generate *m* imputed data sets from a distribution which results in *m* complete data sets. The distribution can be different for each missing entry.
- Analysis: In this stage each *m* imputed data Sets the analysis is performed, it is known as complete data analysis.
- Pooling: Use simple rules the output obtained after data analysis is pooled to get final result.

The resulting inferences form this stage is statistically valid if the methods to create imputations are "decent."

For substituting missing values with possible solutions, the multiple imputation method is used. The missing data set is transformed into complete data set using suitable imputation methods that can then be analyzed by any standard analysis method.

Therefore, multiple imputations have become popular in the handling of missing data. In this method, the process is repeated multiple times for all variables having missing values as the name indicates and then analyzed to combine m number of imputed data set into one imputed data set [7], [11].

4. MISSING VALUE IMPUTATION TECHNIQUE

4.1. Mean Imputation

Using this technique, calculate the mean of missing value through using the corresponding attribute value. This technique is easy to apply; it is built in maximum of the statistical bundle and quicker comparing with other techniques. It introduces precise result when facts is small, but it provides not proper result for large facts, this technique is appropriate for only MAR but no longer beneficial for MCAR [8], [13].

$$\hat{x}_{ij} = \sum_{i:x_{ij} \in c_k} \frac{x_{ij}}{n_k}$$

Wherein n_k represents the number of non-missing values within the j-th feature of the *k-th* class C_k, is missing [7], [8].

4.2. Hot (Cold) Deck Imputation

The concept, in this case, is to use some criteria of similarity to cluster the data earlier than executing the data imputation. This is one of the most used strategies.

Hot deck strategies impute missing values inside a data matrix by way of the usage of available values from the equal matrix. The item, from which these available values are taken for imputation within some other, is referred to as the donor. The replication of values ends in the trouble, that a single donor might be selected to accommodate multiple recipients. The inherent risk posed through that is that too many, or even all, missing values can be imputed with the values from a single donor. To mitigate this chance, a few hot deck variants restrict the amount of times anyone donor may be selected for donating its values. The similar techniques of hot deck are cold deck imputation method which takes other data source than current dataset. Using hot deck, the missing values are imputed by realistically obtained values which avoids distortion in distribution, but bit empirical work for accuracy estimation, creates problem if any other sample has no close relation in entire manner of the dataset [8], [10], [11].

4.3. Median Imputation (MDI)

Due to the affected of the mean through the presence of outliers, it seems better to use the median rather simply to make certain robustness. In this situation, the missing data are changed through the median of all recognized values of that attribute within the class where the instance with the missing characteristic belongs. This method is likewise a considered as a choice whilst the distribution of the values of is skewed. Assume that the value x_{ij} of the *k-th* class, C_k, is missing. It will get replaced by means of Singh and Prasad [7].

$$\hat{u}\ddot{y}_{ij}\ddot{u} = \quad \left(i{:}x_{ij} \in c_k\right)\left\{x_{ij}\right\}$$

4.4. Regression Imputation

This approach may be apply by the use of known values for the construction of model after which calculates the regression between variables ends with applying that technique to calculate the missing values. The outcomes from applying this technique give greater accurate than mean imputation. The calculated data saves deviations from mean and distribution shape but the degree of freedom gets distort and can increases relationship [10].

$$Y = \alpha 0 + \alpha 1 \ X$$

4.5. Expectation Maximization Imputation (EMI)

There are forms of clustering algorithms. One is soft clustering and other is hard clustering:-

- *Soft clustering:* Clusters may overlap that is with unique degree of belief the factors belong to multiple clusters at the identical time
- *Hard clustering:* Clusters do now not overlap that's mean the element either belong to a cluster or not.
- *Mixture models:* The use of a probabilistic manner for doing soft clustering. Every cluster corresponds to a generative model this is usually Gaussian or multinomial, MVs are imputed by realistically obtained values which avoids distortion in distribution, in this technique, bit empirical work for accuracy estimation creates problem if any other sample has no close relation in entire manner of the dataset [2].

4.6. K-nearest Neighbor Imputation (KNN)

Specifying the similarity between two values and replace the missing value with similar one using Euclidean distance. The advantage of this technique that for the datasets which having both qualitative and quantitative attributes values KNN is suitable. There is no need for creating a predictive model for each attribute of missing data and helpful for multiple missing values.

The KNN looks for the most similar instances, the algorithm searches through all of the data set and that consider as an obstacle for that approach [12].

4.7. Fuzzy K-means Clustering Imputation (FKMI)

In this method, the membership characteristic plays an important position. It is assigned with every data item that describes in what degree the data object is belonging to the precise cluster. data items might not get assign to concrete cluster which is stated using centroid of cluster (i.e., the case of k means), that is due to the various membership degrees of every data with entire k clusters. Unreferenced attributes for every uncompleted data are changing by FKMI on the premise of membership degrees and cluster centroid values. The pros of this approach is that it offers quality outcome for overlapping data, higher than k manner imputation and records objects may be a part of multiple cluster middle but the high computation time and noise sensitive, that is, low or no membership degree for noisy objects considered as a cones for the usage of this technique [10].

4.8. Support Vector Machine Imputation (SVMI)

Its regression primarily based technique to impute the missing values. It takes condition attributes (output) and decision attributes. Then, the SVMI would be carried out for prediction of values of missed condition features. Advantages of this approach are the efficient in massive dimensional areas and efficient memory consumption; however, additionally, there may be a cons for using this technique which it is the bad performance if number of samples are plenty lesser than number of features [10], [14].

4.9. Most Common Imputation (MCI)

On this imputation method, clustered are first shaped by applying k-means clustering method. Like in k-NN, on this method, the nearest neighbors are found using clusters. All the instances in every cluster are referred as nearest neighbor of each other. Then, the missing value is imputed the usage of the same technique as is employed through KNNI imputation approach. This procedure is fast and therefore is ideal for applying in big datasets. This algorithm reduces the intra cluster variance to minimum. Here, too value of k parameter is an important factor and is difficult to predict its value. In addition, this algorithm does no longer assure global minimal variance [2], [15], [16].

4.10. Multivariate Imputation by Chained Equation (MICE)

MICE expect that data are lost arbitrarily (damage). It imagines the likelihood of a missing variable depends on the watched facts. MICE offers numerous values in the put of one lost esteem through making an arrangement of relapse (or other reasonable) models, tallying on its “method” parameter. In MICE, every lost variable is treated as a variable, and other information inside the record is treated as an independent variable. At to begin with, MICE foresee missing values utilizing the winning information of other factors. At that point, it replaces missing values utilizing the predicted values and makes a dataset known as ascribed dataset. By cycle, it makes numerous ascribed datasets. Every dataset is at that factor analyzed utilizing standard measurable investigation techniques, and numerous investigation comes about are given [17], [18].

4.11. SICE Technique

It pretends the probability of a missing variable depends on the determined data. It gives multiple values within the place of one missing value through creating a sequence of regression models, each missing variable is treated as a dependent variable, and different data in the record are treated as an independent variable, it predicts missing data using the existing data of other variables. Then, it replaces missing values using the predicted values and creates a dataset known as imputed dataset. It achieves 20% higher F-measure for binary data imputation and 11% less errors for numeric data imputations than its competitors with similar execution time. It imputes binary, ordinal and numeric data. It performed well for the imputation of binary and numeric data and fantastic preference for missing data imputation, especially for massive datasets where MICE is impractical to use because of its complexity but it could not show better overall performance than MICE for the case of ordinal data [6].

4.12. Reinforcement Programming

Impute missing data using learning a policy to impute data thru an action-reward-based totally experience imputes missing values in a column by operating best on the identical column (similar to univarite single imputation) however imputes the missing values within the column with different values thus keeping the variance in the imputed values. It is usually used for dynamic approach for the calculation of missing values using machine learning procedures. It has functionality of convergence and to solving imputation problem through using exploration and exploitation [19], [20].

4.13. Utilizing Uncertainty Aware Predictors And Adversarial Learning MIP UA-Adv.

Impute the missing values so that the adversarial neural network cannot distinguish real values from imputed ones. In addition, to account for the uncertainty of imputed values, the usage of confidence scores acquired from the adversarial module. The adversarial module objectives to discriminate imputed values from real ones the resulting imputer in addition to estimating a missing entry with high accuracy, it

Table 1: Short review with mentioning to the advantage and disadvantage of different techniques to handle missing value

Techniques	Note	Advantages	Limitations	References
Leastwise deletion technique	- Deletion of cases containing missing values (complete row is deleted) high missing information because of deletion of entire row high impact on variability loss of precision and induce bias.	- Simple to use.	- Loss of precision, - Loss of enormous data - High effect on variability, - Induce bias	[2], [10]
Pair- wise deletion technique	- Deletion of records best from column containing missing values much less lack of information by using keeping all available values less impact on variability less loss of precision and induce bias.	- Keeping all available values only missing values are deleted. - Simple to use.	- Not a better solution as compared to other methods. - Loss of data,	[2], [10]
Mean Imputation technique	- Calculate the mean of missing value through using the corresponding attribute value. Replace MVs with the mean of facts Resultant may be better than that of original.	- It is built in maximum of the statistical bundle and quicker comparing with other techniques. - It introduce precise result when facts is small	- It provides not proper result for large facts this technique is appropriate for only MAR but no longer beneficial for MCAR - Affected by the presence of outliers.	[3], [8]
Median imputation (MDI) technique	- Missing data replaced by the median of all observed values of that attribute in the class where the features belongs.	- Good choice when the distribution of the values is skewed.	- Not affect by the presence of outlier	[7]
Hot (cold) deck imputation technique	- Cluster the data earlier than executing the data imputation. - Impute missing values inside a data matrix by way of the usage of available values from the equal matrix	- Avoid distortion in distribution.	- Empirical for accuracy estimation. - creates problem if any other sample has no close relation in entire manner of the dataset.	[8], [10], [11]
Regression imputation technique	- Use the known values for the construction between variables then applying the technique to calculate the missing values	- Very easy and simple technique. - Calculated data saves deviations from mean and distribution shape	- Only applicable if data is linearly separable that is there is linear relation between attributes. - Degree of freedom gets distort and may raises relationship.	[10]
Expectation maximization (EM) technique	- Iterative method, finds maximum likelihood Two steps: Expectation (E step), Maximization (M step) using three models soft, hard and mixture clustering Iteration goes on until algorithm converges	- MVs are imputed by realistically obtained values which avoids distortion in distribution	- Bit empirical work for accuracy estimation, creates problem if any other sample has no close relation in entire manner of the dataset	[2]
Fuzzy K- means clustering Imputation (FKMI) technique	- It is assigned with every data item that describes in what degree the data object is belonging to the precise cluster. - Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values.	- Best outcome for overlapping data, better than k means imputation. Data objects may be part of more than one cluster center	- High computation time. - Noise sensitive, that is, low or no membership degree for noisy objects	[10]
- Support Vector Machine Imputation (SVMI) technique	- Takes condition attributes (here, decision attribute i.e., output) and decision attributes (here, conditional attributes) SVMI then would be applied for prediction of values of missed condition attribute	- Efficient in large dimensional spaces. - Efficient memory consumption	- Poor performance if number of samples are much less than number of feature	[10], [14]

(Contd...)

Table 1: (*Continued*)

Techniques	Note	Advantages	Limitations	References
K nearest neighbour imputation (KNN) technique	- Determining the similarity between two values and replace the missing data with similar one using Euclidean.	- Avoids distortion in distribution as missing values are imputed by realistically obtained values - No need for creating a predictive model. - Helpful for multiple missing value	- Obstacle approach since the algorithm search all of the data set - Prediction of value of k is quite a difficult task.	[12]
Most Common Imputation (MCI) technique	- It replaces the missing value by the most common attribute or by the mode. - While the numerical attribute missing value replaced by the average of the mean corresponding attribute	- Fast and good for applying in big dataset. - Reduce the intra cluster variance to minimum.	- Difficult to predict the value if the number of elements too big. - Dose not guarantee global minimum variance.	[2], [15], [16]
Multivariate Imputation by Chained Equation (MICE) technique	- It pretends the probability of a missing variable depends on the observed data. it provides multiple values in the place of one missing value by creating a series of regression models, - Each missing variable is treated as a dependent variable, and other data in the record are treated as an independent variable - Predict missing data using the existing data of other variables. Then it replaces missing values using the predicted values and creates a dataset called imputed dataset	- Flexibility: each variable can be modeled using a model tailored to its distribution. - Can manage imputation of variables defined only on a subset of the data, - Can also incorporate variables that are functions of other variables, - It does not require monotone missing- data patterns.	- Lacking a theoretical rationale - Difficulties encountered when specifying the different imputation models	[17], [18]
SICE technique:	It is an extension of the popular MICE algorithm. Two variants of SICE presented: SICE- Categorical and SICE- Numeric to impute binary, ordinal, and numeric data. Twelve existing Performance of algorithms implemented to predict house prices imputation methods and compare their performance with SICE.	- Achieves 20% higher F- measure for binary data imputation and 11% less error for numeric data imputations than its competitors with similar execution time. Impute binary, ordinal, and numeric data. - Performed better for the imputation of binary and numeric data. - Excellent choice for missing data imputation, especially for massive datasets where MICE is impractical to use because of its complexity	- It could not show better performance than MICE for the case of ordinal data.	[6]
Reinforcement programming technique	Impute data through an action-reward- based experience imputes missing values in a column by working only on the same column but imputes the missing values in the column with different values thus keeping the variance in the imputed values. It is generally used for dynamic approach for the calculation of missing values by using machine learning approaches.	- Performs well compared to other univarite single imputation and ML- based imputation approaches.	- Use of numeric data variables only	[19], [20]

(Contd...)

Table 1: (*Continued*)

Techniques	Note	Advantages	Limitations	References
Utilizing uncertainty aware predictors and adversarial learning MLP UA- Adv Imputer	- Train well with small and large datasets and utilizes a novel adversarial strategy to estimate the uncertainty of imputed data - Proposed a novel hybrid loss function that enforces the imputers to generate values for missing data that on the one hand, obey the underlying data distribution so that it can confuse the well- trained adversarial module, and on the other hand, predict existing non- missing values accurately - The run time of the methods shows that they are efficient and have less execution time in comparison with that of peer imputer models.	- Plays an important role in the overall performance - Less runtime compared to other imputers - Has a very simple structure, can work with any feature type and small and large data set	- It did not consider the imbalanced nature of the imputation task.	[19], [21]

Table 2: Comparing different techniques according to the dataset used in the application

Datasets	Techniques	Notes	References
Iris	Mean Regression Imputation; Reinforcement Programming technique	A comparison of different approaches of MICE methods on iris datasets. Efficiency gain with multiple imputations combined with regression is that it can better use the available information by accommodating non- linarites	[3], [8], [10], [18], [19], [20]

Iris
Credits
Adults
	Mean/Mode;
	Hot Deck;
	Expectation Maximization;
	K- nearest neighbor
		In this paper, the authors compare C5.0 with this newly developed technique known as IITMV and show its performance on different data sets	[3], [8], [10], [11], [12], [22]

Cleveland Heart
Zoo
Buhl1- 300
Glass
Ionosphere
Iris
Pima
Sonar
WaveForm2
Wine
Hayes- Roth
Led7
Solar
Soybean
	Mean/mode;
	Regression;
	Hot deck;
	MLP UA- Adv
		The result shows that multilayer perceptions (MLP) with different learning rules show better results with quantitative datasets than classical imputation methods. In this paper, the type of missing value is missing completely at random (MCAR)	[3], [8], [10], [11], [19], [21], [22]

Iris *Escherichia coli* Breast cancer 1 Breast cancer 2	Mean K- nearest neighbors (KNN) Fuzzy K- means (FKM) Multiple imputations by chained equations (MICE) MLP UA- Adv	The results show that different techniques are best for different datasets and sizes. MICE are useful for small datasets, but, for big ones and FKM are better, the MLP UA- Adv is better for both small and big datasets	[3], [8], [10], [12], [17], [18], [19], [21], [23]

be able to confuse the adversarial module, it neural network based totally architecture that can train properly with small and large datasets and to estimate the uncertainty of imputed data [19], [21].

5. REVIEW ON MISSING VALUE IMPUTATION METHODS

Table 1.

6. CONCLUSION

The finding of this article summarized in Tables 1 and 2, the article shows that the most popular techniques (mean, KNN, and MICE) are not necessarily the most efficient. It isn't always surprising for mean in regards to the simplicity of the method: The technique does not make use of the underlying correlation structure of the information and for that reason plays poorly. KNN represents a natural improvement of mean that exploits the observed facts structure. MICE are complex algorithm and its behavior seems to be related to the size of the dataset: Rapid and efficient on the small datasets, its overall performance decreases and it becomes time-intensive when carried out to the massive datasets. The more than one imputation combined with Bayesian Regression gives better performance than other strategies, which includes mean, KNN. However, they only taken into consideration the great of imputation based totally on category strategies without worrying of the execution time that may be an exclude criterion. Consequently, FKM may additionally represent the technique of choice but its execution time may be a drag to its use and we take into account bPCA as a more adapted solution to high-dimensional data, the article also shows that the MLP UA-Adv consider a good choice for large and small data set also with different data type. Table 2 shows comparison between the techniques according applications and the dataset used in each one. The strength of this paper that its cover most of the missing value imputation techniques that can be taken into consideration as a reference for other researcher to pick out the most appropriate techniques or make combination from a couple of for imputing the missing values.

REFERENCES

[1] B. Doshi. Handling Missing Values in Data Mining. Rochester Institute of Technology, Rochester, New York, U S A, 2010. Available from: https://www.pdfs.semanticscholar.org/3817/b208fe1f40891cc661ea0db80c8fccc56b70.pdf [Last accessed on 2023 Mar 27].

[2] S. Gupta and M. K. Gupta. "A survey on different techniques for handling missing values in dataset". *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 4, no. 1, pp. 2456-3307, 2018.

[3] A. Jadhav, D. Pramod and K. Ramanathan. "Comparison of performance of data imputation methods for numeric dataset". *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913-933, 2019.

[4] J. Scheffer. "Dealing with missing data". *Research Letters in the Information and Mathematical Sciences*, vol. 3, pp. 153-160, 2002.

[5] D. V. Patil. "Multiple imputation of missing data with genetic algorithm based techniques". *IJCA Special Issue on Evolutionary Computation*, vol. 2, pp. 74-78, 2010.

[6] S. I. Khan and A. S. Hoque. "SICE: An improved missing data imputation technique." *Journal of Big Data*, vol. 7, no. 1, p. 37, 2020.

[7] S. Singh and J. Prasad. "Estimation of missing values in the data mining and comparison of imputation methods." *Mathematical Journal of Interdisciplinary Sciences*, vol. 1, no. 2, pp. 75-90, 2013.

[8] I. Pratama, A. E. Permanasari, I. Ardiyanto and R. Indrayani. A Review of Missing Values Handling Methods on Time Series Data, in: International Conference on Information Technology Systems and Innovation (ICITSI). Bandung, Bali, IEEE, 2016, p. 6.

[9] S. Wang and H. Wang. Mining Data Quality in Completeness. University of Massachusetts Dartmouth, United States of America, 2007. Available from: https://www.pdfs.semanticscholar.org/347c/f73908217751c8d5c617ae964fdcb87674c3.pdf [Last accessed on 2023 Mar 27].

[10] R. L. Vaishnav and K. M. Patel. "Analysis of various techniques to handling missing value in dataset". *International Journal of Innovative and Emerging Research in Engineering*, vol. 2, no. 2, pp. 191-195, 2015.

[11] A. Raghunath. Survey Sampling Theory and Applications. Academic Press, Cambridge, 2017.

[12] Holman and C. A. Glas. "Modelling non-ignorable missing-data mechanisms with item response theory models". *British Journal of Mathematical and Statistical Psychology*, vol. 58, no. 1, pp. 1-17, 2005.

[13] A. Puri and M. Gupta. "Review on missing value imputation techniques in data mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, no. 7, pp. 35-40, 2017.

[14] S. Van Buuren and K. Groothuis-Oudshoorn. "MICE: Multivariate imputation by chained equations in R". *Journal of Statistical Software*, vol. 45, no. 3, pp. 1-67, 2010.

[15] A. S. Kumar and G. V. Akrishna. "Internet of things based clinical decision support system using data mining techniques". *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 4, pp. 132-139, 2018.

[16] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse and X. Zheng. Handling Missing Attribute Values in Preterm Birth Data Sets. Vol. 3642. United Nations Academic Impact, New York, 2005, pp. 342-351.

[17] J. Han, M. Kamber and J. Pei. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2012.

[18] G. Chhabra, V. Vashisht and J. Ranjan. "A comparison of multiple imputation methods for data with missing values". *Indian Journal of Science and Technology*, vol. 10, no. 19, pp. 1-7, 2017.

[19] S. E. Awan, M. Bennamoun, F. Sohel, F. Sanfilippo and G. Dwivedi. "A reinforcement learning-based approach for imputing missing data". *Neural Computing and Applications*, vol. 34, pp. 9701-9716, 2022.

[20] I. E. W. Rachmawan and A. R. Barakbah. Optimization of Missing Value Imputation using Reinforcement Programming, in:

International Electronics Symposium (IES). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 2015, pp. 128-133.

[21] W. M. Hameed and N. A. Ali. “Enhancing imputation techniques performance utilizing uncertainty aware predictors and adversarial learning”. *Periodicals of Engineering and Natural Sciences*, vol. 10, no. 3, pp. 350-367, 2022.

[22] T. Aljuaid and S. Sasi. Intelligent Imputation Technique for Missing Values, in: Conference on Advances in Computing, Communications and Informatics (ICACCI). Jaipur, India, pp. 2441-2445, 2016.

[23] P. Schmitt, J. Mandel and M. Guedj. “A comparison of six methods for missing data imputation”. *Journal of Biometrics and Biostatistics*, vol. 6, no. 1, pp. 1, 2015.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Exploring the Relationship between Attitudes and Blood Glucose Control among Patients with Type 2 Diabetes Mellitus in Chamchamal Town, Kurdistan, Iraq

Hawar Mardan Mohammed[1]*, Samir Yonis Lafi[1]

[1]Department of Nursing, University of Raparin, Kurdistan Regional Government, Iraq, [2]Department of Nursing, College of Nursing, University of Human Development, Kurdistan Regional Government, Iraq

ABSTRACT

Background: Diabetes mellitus type 2 is an endocrine disorder characterized by a progressive elevation in blood glucose levels. It is a persistent and incapacitating illness that may result in mortality if not properly managed. **Objectives:** The objective of this research is to explore the relationship between the attitudes of individuals with type 2 diabetes mellitus and their ability to regulate blood glucose levels. In particular, the study aims to investigate the potential correlation between participants' attitudes and their capacity to manage blood glucose levels following their participation in an educational program. Moreover, the research seeks to analyze the association between individuals' attitudes and diabetes control. Ultimately, the study intends to evaluate the levels of participants' attitudes through appropriate measures. **Materials and Methods:** The study is designed as a cross-sectional investigation and utilizes data from a diabetic outpatient center in Chamchamal. The study population consists of outpatients from the evening public clinic and chronic disease control center. Participants are required to complete questionnaires on their diabetes attitude. The study was conducted between August 11, 2019, and January 5, 2022. To explore the efficacy of the attitude with diabetes control, we used a correlation coefficient test and a t-test with *P*-value of 0.05 as our alpha level of significance. **Results and Conclusion:** The study found that the majority of patients with type 2 diabetes mellitus had low levels of educational attainment, were married and had insufficient monthly income. In addition, 85% of the patients reported not smoking, and 48.3% were classified as overweight. These findings highlight the need for health-care providers to consider sociodemographic factors in the management of diabetes mellitus.

Index Terms: Attitude, Type 2 diabetes mellitus, Blood glucose control, Diabetic complications, Self-care management

1. INTRODUCTION

Diabetes is a major public health issue, projected to be the seventh leading cause of death by 2030 [3]. Patients with T2DM patients with suboptimal glycemic control and HbA1c levels are more likely to develop microvascular problems and cardiovascular disease [1,13]. HbA1c levels have been shown to be affected by modifiable psychosocial variables such self-care habits and attitude [2,5]. Without good self-care practices, it might be difficult to keep HbA1c levels in check [3,6] frequency, population distribution. The authors express concern that diabetes might develop into a regional public health problem and suggest measures to combat the disease [4,5]. Developing healthy self-care habits is essential for managing HbA1c levels, which can increase without proper

Access this article online

DOI:10.21928/uhdjst.v7n1y2023.pp82-91

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: Hawar Mardan Mohammed, Department of Nursing, University of Raparin, Kurdistan Regional Government, Iraq. E-mail: hawar.mardan84@gmail.com

Received: 28-11-2022 Accepted: 11-03-2023 Published: 12-04-2023

self-care [6]. Attitude is the degree to which a person believes he or she is capable of doing a job, and attitudes precede actions [7,8]. The amount of self-assurance that individual possesses regarding their capacity to carry out a task [15]. is referred to as their attitude, and it is normal for an individual's attitudes to come before their actions. Patients with diabetes need to make lifestyle changes to manage their blood glucose levels [9]. This study aims to investigate the potential correlation between attitudes of individuals with type 2 diabetes mellitus and their ability to regulate blood glucose levels [10]. patients with diabetes can also improve their health and prevent further complications by losing weight and lowering their body mass index (BMI) [11]. Diabetes attitude is a patient's attitude toward managing the disease, controlling blood sugar, reducing complications, and preventing short-term problems [12]. Effective patient attitude management strategies can reduce the risk of chronic complications and prevent acute complications in type 2 diabetes [10]. Individuals who maintain optimal glycemic control are at a reduced risk of developing microvascular complications, such as those that affect the kidneys, nerves, and eyes. These complications can manifest in the form of cataracts, glaucoma, renal failure, and lower limb amputations. Conversely, when blood glucose levels are maintained at appropriate levels, macrovascular complications, including heart attacks and strokes, appear to be averted [14].

This study aims to investigate the relationship between attitude and ability to manage blood glucose.

2. METHODOLOGY

2.1. Study Design

Sixty patients were studied in this cross-sectional study from the Diabetes and Chronic Disease Control Center in the Chamchamal District of Sulaimaniyah, Iraq, between July 7, 2020, and November 7, 2020.

2.2. Sample Size

Raosoft's sample size calculator was used to determine the appropriate sample size. Only 60 patients out of a possible 2000 at the Diabetes and Chronic Disease Control Center were included in this research.

2.3. Inclusion Criteria

The research study exclusively included adult patients who had been diagnosed with type 2 diabetes and met the rigorous eligibility criteria set forth by the trial. To be included in the study, participants were required to provide informed consent and meet all the necessary prerequisites for research participation. The eligible individuals who met the inclusion criteria are described in detail below.

2.4. Exclusion Criteria

Patients with T1DM, pregnant women with T2DM, liver failure, impairments or special requirements, and gestational diabetes were excluded from the study.

2.5. Ethical Approval

The University approved the moral viewpoints expressed by the Ethics Committee of the College of Nursing at Raparin. In addition, participants were informed of the purpose and nature of the research.

2.6. Patient Informed Consent

Before data were collected, participants were asked to sign informed consent forms and give their verbal and written informed consent in Kurdish. They were also what might come out of the study. Furthermore, a lot of thought goes into patients' rights, privacy, and the safety of their information.

2.7. Questionnaire

A questionnaire to evaluate a patient's attitude and behavior was designed and composed of 3 parts that covered sociodemographic factors, clinical parameters, and attitude behaviors evaluation. Each section uses a Likert scale to rate the respondent's degree of agreement with each statement. The participants' total replies were computed on a scale from 1 to 30, with Always = 1, Sometimes = 2, and Never = 3. Then, the attitude score was determined for each participant based on their responses to sets of 30 questions. The Likert questionnaire had a reliability of 0.92 based on the results of the Cronbach's Alpha test; then the items were presented to all patients in the same order. After taking the patient's height and weight, the BMI (BMI; kg/m^2) was determined. BMI stands for body mass index, and it is a measure of a person's body fat based on their height and weight. It is calculated by dividing a person's weight in kilograms by their height in meters squared (kg/m^2). BMI is a commonly used metric for determining whether an individual's weight is within a healthy range or if they are overweight or obese. It is often used in both clinical and research settings as a quick and easy screening tool for assessing a person's weight status and associated health risks. However, it should be noted that BMI is not a perfect measure and has certain limitations, such as not taking into account body composition or distribution of body fat, a researcher used a targeted sampling technique to obtain data.

2.8. Measure of the Clinical Parameter

The BMI was classified according to the WHO criteria in which <18.5 kg/m^2 = Underweight, 18.5–24.9 kg/m^2 = Normal weight, 25.0–29.9 kg/m^2 = Pre-obesity, 30.0–34.9 kg/m^2 = Obesity Class I, 35.0–39.9 kg/m^2 = Obesity Class II, and <40 kg/m^2 = Obesity Class III World Health Organization (2021).

2.9. Statistical Analysis

SPSS version 25 was utilized for conducting data analysis.

3. RESULTS

3.1. Participants' Demographic and Clinical Characteristics

Patients in this research had a mean age of 58.07 ± 0.309 years, a median age of 57.5 years, and an age range of 39–81 years. Regarding educational attainment, the majority of patients (55%) were illiterate, followed by elementary school graduates (28.3%), and only 1.7% were college graduates or postgraduates. In contrast, 90% of patients were married, compared to 1.7% who were single or separated (not living together). Regarding the patients' employment, the majority (40%) were housewives, whereas the minority (10%) were retirees. The majority of patients have insufficient monthly income (65%), reside in urban areas (73.3%), do not smoke (85%), and are overweight (48.3%). The majority of patients had T2DM for 10 years and took antihyperglycemic therapy orally (98.3%) (Table 1).

3.2. Changes in Attitudes and Practices before and after the Intervention

Table 2 shows the terms of some of the differences between the pre- and post-test attitudes toward controlling disease, the distribution of the mean scores of the pre- and post-test attitudes and practices toward the daily care of patients, and the associated constructs. The table also shows the attitudes and actions that have the most to do with stopping diseases. For example, the highest mean score for the total number of possible points in the pre-attitude group was 2.98 (I eat or drink regularly every day), while the lowest was 1.3 (I try to learn how to control my diabetes by going to different diabetes education programs) (Table 2a). The highest mean score for the total number of possible points in the post-attitude group was also 2.98. The point with the lowest mean score was 1.77, which said that herbal medicines have fewer side effects than medical ones (Table 2b).

3.3. Correlation between Attitudes and Sociodemographic Characteristics

Table 3 presents a correlation matrix that facilitates the examination of the relationship between attitudes (before

TABLE 1: The T2DM patients' (No.=60) sociodemographic and clinical information

Variable	Frequency	Percent
Level of education		
Illiterate	33	55.0
Primary school graduate	17	28.3
Secondary school graduate	7	11.7
Institute graduate	2	3.3
Collage and post graduate	1	1.7
Marital status		
Single	1	1.7
Married	54	90.0
Widow	2	3.3
Divorced	2	3.3
Separated (not living together)	1	1.7
Occupation		
Government employed	9	15.0
Self employed	12	20.0
Retired	6	10.0
House wife	24	40.0
Jobless	9	15.0
Monthly income		
Sufficient	4	6.7
Barely sufficient	17	28.3
Insufficient	39	65.0
Residential area		
Urban	44	73.3
Rural	16	26.7
Duration of diabetes mellitus		
≤10 years	45	75.0
≤20 years	12	20.0
>20 years	3	5.0
Treatment method		
Oral antihyperglycemic agents	59	98.3
Insulin	1	1.7
Do you smoke?		
Yes	9	15.0
No	51	85.0
How many cigarettes per day?		
11–20	1	10
21–30	9	90
Body mass index		
Underweight	1	1.7
Normal weight	10	16.7
Over weight	29	48.3
Obesity I	14	23.3
Obesity II	5	8.3
Obesity III	1	1.7
For how many years have you smoked?		
10	2	15.38
15	5	38.46
20	2	15.38
25	3	23.08
40	1	7.69
Source of information about disease		
Physician	40	66.7
Nurse	13	21.7
Books and magazines	1	1.7
Television	6	10.0

TABLE 2a: The participants pre-attitude behaviors evaluation

Variable	Always=3	Sometime=2	Never=1	Mean score	Rank	%
I visit hospital regularly according to doctor's appointment for examination or treatment of diabetes.	26	25	9	2.28	2	64
I take meals or refreshment regularly every day.	14	40	6	2.13	8	56.5
I eat as well-balance diet using a list of food exchanges	15	42	3	2.2	5	60
I take foods containing dietary fiber like grain, vegetable and fruit every day.	14	41	5	2.15	7	57.5
I set a limit of taking salt and processed foods.	29	20	11	2.3	1	65
I do a self-blood sugar test according doctors' recommendations.	14	30	16	1.97	10	48.5
I do a self-blood sugar test more frequently, when I feel symptoms of hypoglycemia such as tremor, pallor, and headache.	14	24	22	1.87	12	43.5
I try to maintain the optimal blood sugar level.	9	34	17	1.87	13	43.5
I control the size of meals or exercise according to a blood sugar level.	6	32	22	1.73	17	36.5
I am carrying food likes sweet drink, candy or chocolate just in case of hypoglycemia.	3	20	37	1.43	25	21.5
I try to maintain optimal weight by measuring my weight regularly.	5	32	23	1.7	18	35
I carry insulin, injection and blood sugar tester whenever I go to trip.	5	10	45	1.33	27	16.5
I try to get information on diabetes control by attending various diabetes educational programs.	4	10	46	1.3	30	15
I take my diabetes medication like insulin injection as prescribe observing dosage and time regularly.	6	14	40	1.43	26	21.5
I keep in touch with my physician.	16	41	3	2.22	4	61
Herbal medications have less complications than medical medications	3	27	30	1.55	20	27.5
Regular exercise helps me to control diabetes.	5	23	32	1.55	21	27.5
Reading handouts on proper footwear is necessary for me.	3	13	44	1.32	29	16
Blood pressure control helps me to control my diabetes mellitus.	8	38	14	1.9	11	45
Annual eyes examination is necessary for me.	9	34	17	1.87	14	43.5
Always I be relaxing and avoid stress and bad mood because its effects diabetes negatively.	10	43	7	2.05	9	52.5
I did not miss doses of my diabetic medication.	20	31	9	2.18	6	59
I inspect my feet during and after my shower/bath.	14	21	25	1.82	15	41
I use talcum powder to keep my inter-digital spaces dry.	6	20	34	1.53	22	26.5
I check the temperature of water before use.	2	16	42	1.33	28	16.5
I examine my feet daily.	6	17	37	1.48	23	24
I used to check my blood glucose level.	5	27	28	1.62	19	31
I used to check fasting blood glucose and 2 h after meal by glucometer.	6	37	17	1.82	16	41
I take my medication according of physician recommendation.	24	27	9	2.25	3	62.5
I did not wear tide shoes.	8	13	39	1.48	24	24

and after) and sociodemographic characteristics. The matrix displays only those variables that exhibit a statistically significant correlation ($P < 0.05$) as determined by Pearson's r, with regard to satisfaction levels of the simulation experience. The matrix allows for an assessment of the degree of association between sociodemographic factors and participant satisfaction with the simulation, providing valuable insights into the factors that may impact user experience. The Mann–Whitney U-test is used in.

3.4. Gender Differences in Attitudes before and after the Intervention

Table 4 to compare the means of attitudes (before and after) by gender. Before the test, the mean attitude score for males

TABLE 2b: The participant's post-attitude behaviors evaluation

Variable	Always=3	Sometime=2	Never=1	Mean score	Rank	%
I take foods containing dietary fiber like grain, vegetable and fruit every day.	58	2	0	2.97	4	98.5
I set a limit of taking salt and processed foods.	59	1	0	2.98	2	99
I do a self-blood sugar test according doctors' recommendations.	55	5	0	2.92	9	96
I do a self-blood sugar test more frequently, when I feel symptoms of hypoglycemia like tremor, pallor and headache.	47	13	0	2.78	13	89
I try to maintain the optimal blood sugar level.	41	19	0	2.68	15	84
I control the size of meals or exercise according to a blood sugar level.	37	23	0	2.62	21	81
I am carrying food likes sweet drink, candy or chocolate just in case of hypoglycemia.	10	46	4	2.1	29	55
I try to maintain optimal weight by measuring my weight regularly.	29	31	0	2.48	23	74
I carry insulin, injection and blood sugar tester whenever I go to trip.	14	43	3	2.18	28	59
I try to get information on diabetes control by attending various diabetes educational programs.	48	12	0	2.8	12	90
I take my diabetes medication like insulin injection as prescribe observing dosage and time regularly.	43	15	2	2.68	16	84
I keep in touch with my physician.	56	4	0	2.93	8	96.5
Herbal medications have less complications than medical medications.	12	22	26	1.77	30	38.5
Regular exercise helps me to control diabetes.	47	11	2	2.75	14	87.5
Reading handouts on proper footwear is necessary for me.	41	18	1	2.67	18	83.5
Blood pressure control helps me to control my diabetes mellitus.	55	5	0	2.92	10	96
Annual eyes examination is necessary for me.	58	2	0	2.97	5	98.5
Always I be relaxing and avoid stress and bad mood because its effects diabetes negatively.	36	24	0	2.6	22	80
I did not miss doses of my diabetic medication.	38	22	0	2.63	19	81.5
I inspect my feet during and after my shower/bath.	28	32	0	2.47	24	73.5
I use talcum powder to keep my inter-digital spaces dry.	17	43	0	2.28	27	64
I check the temperature of water before use.	28	32	0	2.47	25	73.5
I examine my feet daily.	19	40	1	2.3	26	65
I used to check my blood glucose level.	41	19	0	2.68	17	84
I used to check fasting blood glucose and 2 h after meal by glucometer.	38	22	0	2.63	20	81.5
I take my medication according of physician recommendation.	52	8	0	2.87	11	93.5
I did not wear tide shoes.	58	2	0	2.97	6	98.5

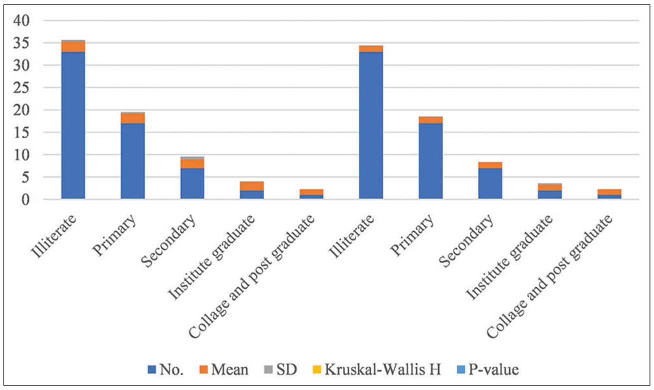


Fig. 1. Compare means of attitude (pre and post) by level of education using Kruskal–Wallis H-test.

was 2.219 and for females it was 2.201. After the test, the mean attitude score for males was 1.353 and for females it was 1.308. Neither score changed significantly from the pre-test. The results of the post-test showed that there wasn't a big difference between men and women using the Kruskal–Wallis H-tes.

3.5. Impact of Education Level on Attitudes before and after the Intervention

Fig. 1 shows the study compared the mean attitudes of participants before and after the intervention with respect to their level of education. The results indicated that there was no statistically significant difference between the mean pre-test and post-test attitudes of the participants.

TABLE 3: Correlation matrix of attitude (pre and post) with the socio-demographic data

Variable	No.	Pre-attitude		Post-attitude		
		P-value	Spearman rank correlation	No.	*P*-value	Spearman rank correlation
Age	60	0.003	0.378**	60	0.716	−0.048
Level of education	60	0.039	−0.268*	61	0.598	0.069
Family member has diabetes mellitus	60	0.598	−0.069	62	0.009	0.334**
Monthly income	60	0.002	0.384**	63	0.753	0.041
Duration of diabetes mellitus	60	0.712	0.049	64	0.795	0.034
Body mass index (BMI)	60	0.587	−0.072	65	0.739	−0.044
How many cigars per day?	10	0.415	0.291	66	0.107	0.541
For how many years have you smoked?	13	0.601	0.16	67	0.424	0.243
How long ago did you quit smoking?	5	0.581	0.335	68	0.581	0.335

*Significant at 0.001 level, **Significant at 0.05 level.

TABLE 4: Compare means of attitude (pre and post) by gender using Mann-Whitney U-test

Attitude	Gender	No.	Mean	SD	Mann-Whitney U	*P*-value
Mean pre	Male	34	2.219	0.356	424.5	0.794
	Female	26	2.201	0.357		
Mean post	Male	34	1.353	0.143	370.0	0.280
	Female	26	1.308	0.112		

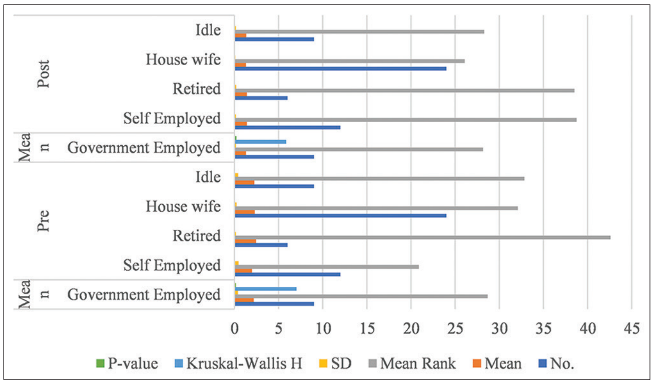


Fig. 2. Compare means of attitude (pre and post) by occupation using Kruskal–Wallis H-test.

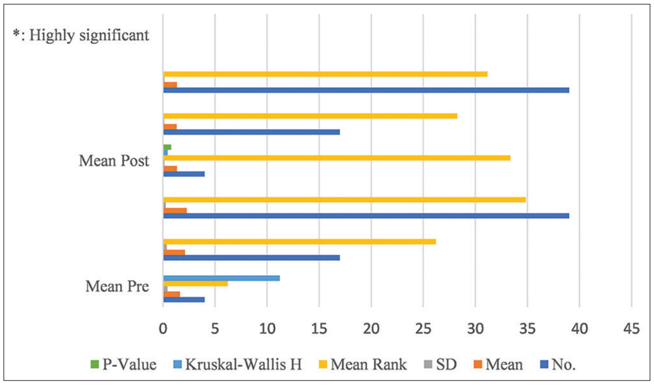


Fig. 3. Compare means of attitude (pre and post) by monthly income using Kruskal–Wallis H-test.

These findings suggest that the intervention did not have a significant impact on the attitudes of participants toward the research topic, regardless of their level of education.

3.6. Impact of Income Level on Attitudes before and after the Intervention

Fig. 2 contrasts the perspectives of patients who participated in the Pattern-Making training before and after which training. It indicates that each group's post-test results for each attitude component differ significantly.

3.7. Statistical Analysis of Income Levels before and after Intervention

In Fig. 3, to assess whether there was a statistically significant difference in the mean rank of income levels, the Kruskal–Wallis test was employed. The results indicated a highly significant difference between the pre-test and post-test means for each income level, with a $P < 0.05$. Specifically, the pre-test mean values ranged from 2.305 to 1.63, while the post-test means ranged from 1.339 to 0.316. The highest pre-test mean was observed as 2.305, while the lowest was 1.63. These findings suggest that the intervention had a

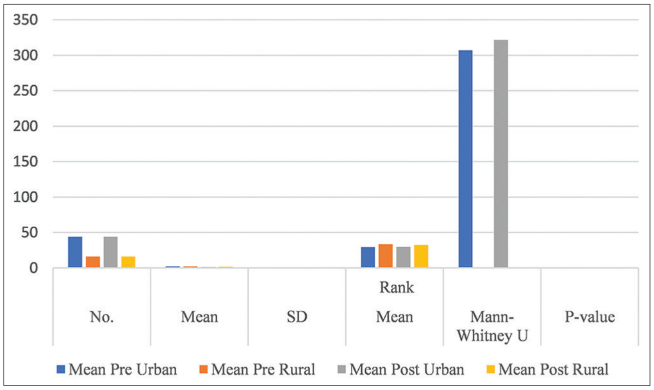


Fig. 4. Compare means of attitude (pre and post) by residential area using Kruskal-Wallis H-test.

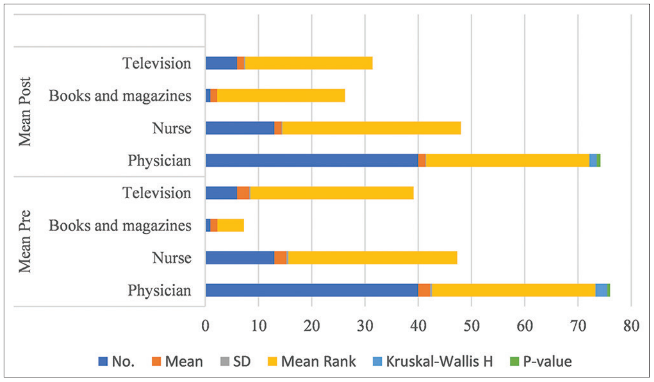


Fig. 5. Compare means of attitude (pre and post) by source of your information about disease using Kruskal–Wallis H-test.

significant impact on the attitudes of participants towards the research topic across all income levels, and support the need for continued efforts to improve attitudes and perceptions. However, to increase the p-value further, it may be necessary to adjust the alpha level or consider a larger sample size.

3.8. Impact of Residential Area on Attitudes before and after the Intervention

In Fig. 4, Mann–Whitney test was conducted to compare the pre- and post-test attitude averages of participants by their residential area. The analysis revealed no significant differences between the pre- and post-test attitudes. The mean attitude score before the intervention was 0.451, while the mean score after the intervention was 0.608.

3.9. Relationship between Source of Knowledge and Attitudes before and after the Intervention

Fig. 5 shows the analysis shows that there is no correlation between the mean attitudes (pre- and post-intervention) and the source of knowledge about the illness. The mean score for the pre-test was 0.529, and for the post-test, it was 0.704, with $P = 0.05$. The average attitude score before the intervention was 0.529, while after the intervention, it was 0.704. The findings indicate that there is no significant relationship between the attitudes of participants before and after receiving information about the disease ($P > 0.05$).

4. DISCUSSION

The median age of T2DM patients in this research was 57.50 years. Consequently, the research population consisted of adults and the elderly. After 50 years of age, the prevalence of (DM) grows progressively, according to the findings of various research conducted in various nations [20]. The majority of our participants were also male. However, national and regional investigations have revealed no substantial gender disparity in the frequency of DM [21]. Therefore, the fact that the majority of our patients were males might be attributed to the fact that men had easier access to hospitals and clinics and more flexible work hours than women.

Before the initiation of the education program, the mean attitude score was found to be similar between the case and control groups. However, after the program was implemented, significant differences were observed between the two groups on multiple attitude-related questions. These findings are consistent with earlier research that employed alternative intervention methodologies, as reported by [23].

In addition, substantial patients completed diabetes self-management education in the current research (DSME). Consequently, they were more likely to comply with the recommended diabetic care standards and their pharmaceutical treatment regimens. This result is consistent with earlier research demonstrating that the Hba1c levels of patients fell considerably following diabetes education program treatments [16]. In addition, another study comparing the opinions of 252 health professionals and 279 individuals with diabetes revealed major disparities in their perspectives. Both groups agreed on the severity of T2DM, the necessity for strict glycemic control, and the psychosocial effect of the condition, but they disagreed on the importance of patient autonomy. This study found no significant differences in the severity of the illness between T1DM and T2DM individuals. In addition, people with diabetes who had previously received diabetes education had elevated rates of disease [17].

There is no significant correlation between gender attitude and program intervention, according to the current study. In contrast, the majority of married couples displayed a level of self-care that fell somewhere in the center. According to Iranian study, married participants in diabetes self-management programs had a more optimistic outlook than their single counterparts [18].

For instance, a separate study found that T2DM patients with stronger marital connections and mutual support have better self-care attitudes and self-management [19]. In addition, there was a substantial correlation between self-care and social support in Iran cross-sectional research [23].

We also demonstrated a substantial relationship between education level and attitude. Patients with higher levels of education, for example, demonstrated more positive attitudes when engaging in diabetic self-management programs. This was especially the case when the programs were implemented. Moreover, illiterates were shown to have a much poorer level of self-care. In addition, it was shown that those with greater levels of education engage in more beneficial kinds of self-care than those with lower levels of education. One study including 125 individuals of diverse racial origins with T2DM found a substantial favorable connection between education and diabetes management [20].

At the outset of the study, the Diabetes Education Program (DEP) evaluated the perspectives of patients with type 2 diabetes on various aspects of the disease, including its physiopathological and nutritional components, treatments, physical activity, patient education, self-monitoring, chronic complications, special situations, and family support. The initial phase of the DEP's development involved assessing the patients' attitude needs toward their illness, followed by an evaluation of their attitudes following the program's implementation. This approach is consistent with the previous studies that have recommended pre- and post-intervention data collection to accurately evaluate the effectiveness of diabetes education programs [24].

The study results suggest that there was a significant change in diabetic patients' perceptions of their illness after participating in the investigation. However, it is difficult to make a conclusive statement about the direct impact of this newfound knowledge on the patients' behavior and lifestyle. While the study revealed that the DEP had a positive effect on the patients' attitudes and behavioral abilities, it was found that the improvements in diet-related attitudes were less significant than those observed for general diabetes knowledge. These findings provide concrete support for the notion that patient education programs can have a positive impact on patients' perceptions of their illness and their ability to manage it effectively. However, further research is needed to determine the specific factors that contribute to behavior change in diabetic patients.

Changing the attitudes of diabetes patients is impacted by a number of factors, including their knowledge of their illness, risk factors, and treatment alternatives. The study investigated the efficiency of group education and determined that it successfully improved and altered attitudes toward self-monitoring of capillary glucose. This was discovered by comparing the attitudes of participants prior to and following the instructional program [25].

This study found a significant association between patient attitude and glycemic control, with patients with more optimistic attitudes exhibiting better glycemic control. This result was supported by a substantial body of literature and contemporary research conducted in several cultural settings [26]. Consequently, according to the American Diabetes Association (ADA), individuals diagnosed with type 2 diabetes mellitus (T2DM) should possess a positive attitude towards their condition to effectively manage the illness and mitigate potential complications. Inadequate glycemic control was associated with a low self-care score, whereas better disease management was associated with a higher self-care score [28].

In the present study, health literacy is shown to significantly influence glycemic control, while higher education levels are associated with favorable health behaviors in patients. Literature supports the notion that health education and literacy can considerably influence illness outcomes, disease management, and prevention of complications. Moreover, patients with higher education levels exhibit more optimistic attitudes compared to those with the lower education levels [27].

Moreover, existing research has established a correlation between health literacy and the mitigation of diabetes complications through the adoption of a positive mindset (Mukanoheli et al., 2020). Furthermore, we observed a notable enhancement in the educational intervention concerning the identification of hypoglycemic symptoms, as inpatient care has been shown to yield better results, a finding that was reinforced during the program's implementation.

5. CONCLUSION

The willingness of patients with type 2 diabetes to adopt a positive attitude and participate in positive behaviors is a significant factor in the effective control of their blood glucose levels in patients with type 2 diabetes. In addition to receiving medical treatment, patients have the additional responsibility of prioritizing healthy daily routines and habits. These may include monitoring their blood glucose levels, making alterations to their food, engaging in physical activity, and caring for their feet. These healthy practices have the potential to have a major influence on the patients' general health and to enhance their capacity to keep their blood glucose levels under control.

6. ACKNOWLEDGMENTS

The Scientific Working Group of the University of Raparin deserves special gratitude for their assistance. The participants and employees at the Center for Diabetes and Chronic Disease Control in the Chamchamal District of Sulaimaniyah, Iraq, also deserve our sincere gratitude.

7. CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] R. M. Anderson and M. M. Funnell. "Patient empowerment: Myths and misconceptions". *Patient Education and Counseling*, vol. 79, no. 3, pp. 277-282, 2010.

[2] M. P. Fransen, C. von Wagner and M. L. Essink-Bot. "Diabetes self-management in patients with low health literacy: Ordering findings from literature in a health literacy framework". *Patient Education and Counseling*, vol. 88, no. 1, pp. 44-53, 2012.

[3] A. G. Brega, A. Ang, W. Vega, L. Jiang, J. Beals, C. M. Mitchell, K. Moore, S. M. Manson, K. J. Acton, Y. Roubideaux and Special Diabetes Program for Indians Healthy Heart Demonstration Project. "Mechanisms underlying the relationship between health literacy and glycemic control in American Indians and Alaska Natives". *Patient Education and Counseling*, vol. 88, no. 1, pp. 61-68, 2012.

[4] G. Danaei, M. M. Finucane, Y. Lu, G. M. Singh, M. J. Cowan, C. J. Paciorek, J. K. Lin, F. Farzadfar, Y. H. Khang, G. A. Stevens, M. Rao, M. K. Ali, L. M. Riley, C. A. Robinson and M. Ezzati. "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: Systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants". *Lancet*, vol. 378, no. 9785, pp. 31-40, 2011.

[5] A. Bener, E. J. Kim, F. Mutlu, A. Eliyan, H. Delghan, E. Nofal, L. Shalabi and N. Wadi. "Burden of diabetes mellitus attributable to demographic levels in Qatar: An emerging public health problem". *Diabetes and Metabolic Syndrome*, vol. 8, no. 4, pp. 216-220, 2014.

[6] World Health Organization. "Diabetes". Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes [Last accessed on 2023 Feb 28].

[7] L. Adam, C. O'Connor and A. C. Garcia, "Evaluating the Impact of Diabetes Self-Management Education Methods on Knowledge, Attitudes and Behaviors of Adult Patients with Type 2 Diabetes Mellitus," Canadian Journal of Diabetes, vol. 42, no. 5, pp. 470-477, 2018.

[8] R. E. Soccio, R. M. Adams, M. J. Romanowski, E. Sehayek, S. K. Burley and J. L. Breslow. "The cholesterol-regulated StarD4 gene encodes a StAR-related lipid transfer protein with two closely related homologues, StarD5 and StarD6". *Proceedings of the National Academy of Sciences U S A*, vol. 99, no. 10, pp. 6943-6948, 2002.

[9] A. van Puffelen, M. Kasteleyn, L. de Vries, M. Rijken, M. Heijmans, G. Nijpels, F. Schellevis and Diacourse Study Group. "Self-care of patients with Type 2 diabetes mellitus over the course of illness: Implications for tailoring support". *Journal Diabetes and Metabolic Disord*ers, vol. 19, no. 1, pp. 81-89, 2020.

[10] L. Mulala. "Diabetes Self-care Behaviors and Social Support among African Americans in San Francisco". Doctoral Dissertation, University of San Francisco; 2017. Available from: https://www.proquestdissertationspublishing [Last accessed on 2023 Mar 02].

[11] V. Mogre, A. Natalie, T. Flora, H. Alix and P. Christine. "Barriers to self-care and their association with poor adherence to self-care behaviours in people with Type 2 diabetes in Ghana: A cross sectional study". *Obesity Medicine*, vol. 18, p. 100222, 2020.

[12] M. A. Powers, J. Bardsley, M. Cypress, P. Duker, M. M. Funnell, A. H. Fischl, M. D. Maryniuk, L. Siminerio and E. Vivian. "Diabetes self-management education and support in Type 2 diabetes: A joint position statement of the American diabetes association, the American association of diabetes educators, and the academy of nutrition and dietetics". *Diabetes Education*, vol. 43, no. 1, pp. 40-53, 2017.

[13] American Diabetes Association. "Classification and diagnosis of diabetes: Standards of medical care in diabetes-2020. *Diabetes Care*, vol. 43, no. Suppl 1, pp. S14-S31, 2020.

[14] F. Moosaie, F. D. Firouzabadi, K. Abouhamzeh, S. Esteghamati, A. Meysamie, S. Rabizadeh, M. Nakhjavani and. A. Esteghamati. "Lp(a) and Apo-lipoproteins as predictors for micro-and macrovascular complications of diabetes: A case-cohort study". *Nutrition Metabolism and Cardiovascular Diseases*, vol. 30, no. 10, pp. 1723-1731, 2020.

[15] M. Baghianimoghadam and A. Ardekani. "The effect of educational intervention on quality of life of diabetic patients Type 2, referee to diabetic research centre of Yazd". *The Horizon of Medical Sciences*, vol. 13, no. 4, pp. 21-28, 2008.

[16] A. Steinsbekk, L. Ø. Rygg, M. Lisulo, M. B. Rise and A. Fretheim. "Group based diabetes self-management education compared to routine treatment for people with Type 2 diabetes mellitus. A systematic review with meta-analysis". *BMC Health Services Research*, vol. 12, no. 1, pp. 213, 2012.

[17] J. J. Gagliardino, C. González and J. E. Caporale. "The diabetes-related attitudes of health care professionals and persons with diabetes in Argentina". *Revista Panamericana de Salud Pública*, vol. 22, no. 5, pp. 304-307, 2007.

[18] M. Reisi, H. Fazeli, M. Mahmoodi and H. Javadzadeh. "Application of the social cognitive theory to predict self-care behavior among Type 2 diabetes patients with limited health literacy". *Journal of*

Health Literacy, vol. 6, no. 2, pp. 21-32, 2021.

[19] J. S. Wooldridge and K. W. Ranby. “Influence of relationship partners on self-efficacy for self-management behaviors among adults with Type 2 diabetes”. *Diabetes Spectrum*, vol. 32, no. 1, pp. 6-15, 2019.

[20] S. S. Bains and L. E. Egede. “Associations between health literacy, diabetes knowledge, self-care behaviors, and glycemic control in a low income population with Type 2 diabetes”. *Diabetes Technology and Therapeutics*, vol. 13, no. 3, pp. 335-341, 2011.

[21] M. Reisi, H. Fazeli, M. Mahmoodi and H. Javadzadeh. “Application of the social cognitive theory to predict self-care behavior among Type 2 diabetes patients with limited health literacy”. *Journal of Health Literacy*, vol. 6, no. 2, pp. 21-32, 2021.

[22] J. S. Wooldridge and K. W. Ranby. “Influence of relationship partners on self-efficacy for self-management behaviors among adults with Type 2 diabetes”. *Diabetes Spectrum*, vol. 32, no. 1, pp. 6-15, 2019.

[23] S. S. Bains and L. E. Egede. “Associations between health literacy, diabetes knowledge, self-care behaviors, and glycemic control in a low income population with Type 2 diabetes”. *Diabetes Technology and Therapeutics*, vol. 13, no. 3, pp. 335-341, 2011.

[24] K. Mulcahy, M. Maryniuk, M. Peeples, M. Peyrot, D. Tomky, T. Weaver and P. Yarborough. “Diabetes self-management education core outcomes measures”. *Diabetes Educator*, vol. 29, no. 5, pp. 768-803, 2003.

[25] M. L. Zanetti, L. M. Otero, M. V. Biaggi, M. A. Santos, D. S. Péres and F. P. de Mattos Guimarães. “Satisfaction of diabetes patients under follow-up in a diabetes education program”. *Revista Latino Americana de Enfermagem*, vol. 15, pp. 583-589, 2007.

[26] C. A. Bukhsh, T. M. Khan, M. S. Nawaz, H. S. Ahmed, K. G. Chan, L. H. Lee and B. H. Goh. “Association of diabetes-related self-care activities with glycemic control of patients with Type 2 diabetes in Pakistan”. *Patient Preference and Adherence*, vol. 12, pp. 2377-2386, 2018.

[27] C. Y. Osborn, S. S. Bains and L. E. Egede. “Health literacy, diabetes self-care, and glycemic control in adults with Type 2 diabetes”. *Diabetes Technology and Therapeutics*, vol. 12, no. 11, pp. 913-919, 2010.

[28] K. Hawthorne, Y. Robles and R. Cannings-John. “Glycemic control and self-care behaviors in Hispanic patients with Type 2 diabetes: A pilot intervention study”. *Journal of Transcultural Nursing*, vol. 23, no. 3, pp. 289-296, 2012.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Boundary Integral Equation Method for Computing Numerical Conformal Mappings onto the Disk with Rectilinear Slit and Spiral Slits Regions

Ali W.K. Sangawi

Department of General Sciences, College of Education, Charmo University, 46023 Chamchamal, Sulaimani, Kurdistan, Iraq.

ABSTRACT

This article proposes a boundary integral equation method for computing numerical conformal mappings of bounded multiply connected region Ω onto the disk with rectilinear slit and spiral slits regions, Ω_1 and Ω_2 Initially, the process involves calculating the boundary value of the canonical region. Cauchy's integral formula can then be used to compute the mapping of the interior values. The effectiveness of the proposed method is demonstrated using several numerical examples.

Index Terms: Numerical conformal mapping, Boundary integral equations, multiply connected regions, Generalized Neumann kernel

1. INTRODUCTION

The identification of canonical regions plays a crucial role in conformal mappings of multiply connected regions. The regions identified as canonical include the disk with circular slits, the annulus with circular slits, the circular slit region, the radial slit region, and the parallel slit region. Furthermore, additional canonical regions for conformal mappings include the disk with spiral slits region, annulus with spiral slits region, spiral slits region, and straight slits region [1]-[13]. Nasser's method of computing conformal mapping is based on Riemann-Hilbert problem [2], [5], [14], while Sangawi's methods rely on integral equations satisfy the interior non-homogeneous boundary relationship [8]-[12].

The canonical slit regions introduced by Koebe [1], DeLillo *et al.* [15], and Nasser [5] are special cases of the spiral slits region. Sangawi [9]-[11] and Sangawi *et al.* [12] have demonstrated conformal mapping of bounded multiply connected regions onto the second, third, and fourth categories of Koebe's canonical slit regions using a boundary integral equation method. In Nasser [14] study, the study of bounded multiply connected region onto the disk with rectilinear slit and spiral slits region was facilitated by reformulating the conformal mapping as a Riemann-Hilbert problem. The present paper aims to establish a new boundary integral equation method for numerical conformal mappings from Ω onto Ω_1 and Ω_2.

The design of the study is as follows: Section 2 presents some necessary materials. A derivation of integral equation method

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp92-99	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Ali W.K. Sangawi, Department of General Sciences, College of Education, Charmo University, 46023 Chamchamal, Sulaimani, Kurdistan, Iraq. E-mail: ali.kareem@charmouniversity.org

Received: 02-04-2023 Accepted: 25-04-2023 Published: 10-06-2023

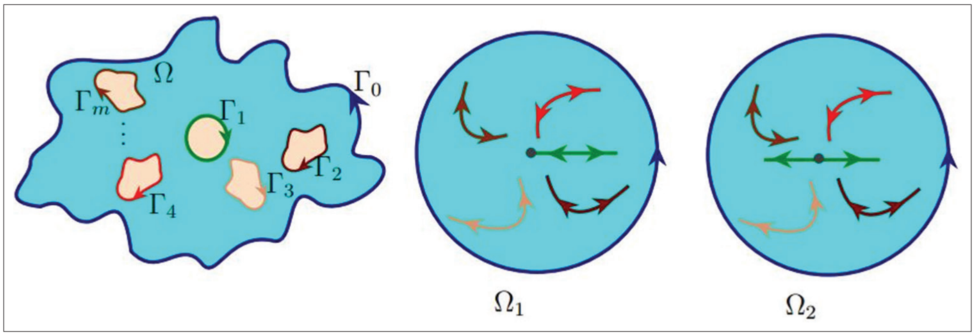


Fig. 1. Mapping of Ω onto Ω_1 and Ω_2.

for computing the function $\mathcal{F}$ has been presented in Section 3. The boundary integral equation method has been illustrated through examples provided in Section 4. Lastly, Section 5 comprises of the conclusion.

2. NECESSARY MATERIALS

A bounded multiply connected region **Ω** of connectivity M + 1. The boundary Γ consists of M + 1 smooth Jordan curves Γ_ι,ι=0,1,…,M as demonstrated in the following, (see Fig. 1)

The curve Γ_ι is parametrized by a 2π periodic twice continuously differentiable complex function $\xi_\iota(t)$

$$\zeta_\iota(t)=\frac{d\,\zeta_\iota(t)}{d\,t}\neq 0\,,\quad t\in_\iota=\left[0\,,2\,\pi\right],\quad \iota=0\,,\ \ldots,M$$

The complete parameter I is the combination of M + 1 disjoint intervals I_ι, ι = 0,…,M. The entire boundary Γ on I is defined by parametrization $\xi(t)$

$$\zeta(t)=\begin{cases}\zeta_0(t), & t\in I_0=\left[0\,,2\,\pi\right],\\ \zeta_1(t), & t\in I_1=\left[0\,,2\,\pi\right],\\ \qquad\vdots & \\ \zeta_m(t), & t\in I_m=\left[0\,,2\,\pi\right]\end{cases}$$

Assuming $\hat{A}(t)$ is a complex function that is continuously differentiable with a periodicity of 2π $\forall t\in I_\iota$. The generalized Neumann kernel that is formed using $\hat{A}$ can be described as[16]:

$$\hat{N}(t\,,s)=\begin{cases}\dfrac{1}{\pi}\operatorname{Im}\left(\dfrac{\hat{A}(t)}{\hat{A}(s)}\dfrac{\zeta'(s)}{\zeta(s)-\zeta(t)}\right), & s\neq t\,,\\[2ex] \dfrac{1}{\pi}\left(\dfrac{1}{2}\operatorname{Im}\dfrac{\zeta''(t)}{\zeta'(s)}-\operatorname{Im}\dfrac{\hat{A}(t)}{\hat{A}(s)}\right), & s=t\,.\end{cases}$$

The classical Neumann kernel is the generalized Neumann kernel formed with $\hat{A}(t)=1$, i.e.

$$N(t\,,s)=\frac{1}{\pi}\operatorname{Im}\left(\frac{\zeta'(s)}{\zeta(s)-\zeta(t)}\right)$$

The adjoint kernel N^* (s,t) of the Neumann kernel is as follows:

$$N^*(t\,,s)=N(s\,,t)=\frac{1}{\pi}\operatorname{Im}\left(\frac{\zeta'(t)}{\zeta(t)-\zeta(s)}\right)$$

The generalized Neumann kernel $\tilde{N}(s\,,t)$ is as follows:

$$\tilde{N}(t\,,s)=\frac{1}{\pi}\operatorname{Im}\left(\frac{\tilde{A}(t)}{\tilde{A}(s)}\frac{\zeta'(s)}{\zeta(s)-\zeta(t)}\right),\tilde{A}(t)=\frac{\zeta'(t)}{\hat{A}(t)}$$

If $\hat{A}=1$, then

$\tilde{N}(t\,,s)=-N^*(t\,,s)$.

Refer to Sangawi [10] for the definitions of $\mathrm{N},\tilde{\mathrm{N}}$ and N^*.

3. INTEGRAL EQUATION METHOD FOR COMPUTING THE FUNCTION $\mathcal{F}$

The canonical region can be described as a disk with a finite straight slit along the line where $\operatorname{Im}\mathcal{F}=0$, as well as M−1 finite spiral slits. Additionally, there is a rectilinear slit, which refers to a slit that lies on a straight line.

$$\operatorname{Im}\left[e^{-i\alpha}\mathcal{F}\right]=r\,,\quad \alpha,r\in\mathbb{R},$$

The variable α represents the angles of intersection between the line and the real axis. There is also a spiral slit, which refers to a slit that is located on a logarithmic spiral.

$$\mathrm{Im}\left[e^{-i\alpha}\log\mathcal{F}\right]=r,\quad \alpha,r\in\mathbb{R},$$

Where the oblique angles α are prescribed in advance.

Assume that the function $\mathcal{F}(\zeta)$ maps the curve Γ_0 onto the circle with radius e^{-R_0}, the circle Γ_1 onto a finite rectilinear slit that lies on the line where Im ($\mathcal{F}(\zeta)$) =0 , and the curves Γ_ι, where ι=2,…,M, onto M−1 finite spiral slits with oblique angles θ_ι,ι=2,…,M. Therefore, the mapping function that transforms Ω onto Ω_1 and Ω_2 fulfills the following conditions.

$$\left|\mathcal{F}(\zeta)\right|=e^{-R_0},\quad t\in I_0 \tag{1}$$

$$\mathrm{Im}\left(\log\left(\mathcal{F}\left(\zeta(t)\right)\right)\right)=0,\quad t\in I_1 \tag{2}$$

$$\mathrm{Im}\left(e^{-i\theta_\iota}\log\left(\mathcal{F}\left(\zeta_\iota(t)\right)\right)\right)=R_\iota,\quad t\in I_\iota,\ \iota=2,\dots,M \tag{3}$$

The values $R_0,R_1,\dots,R_M$ are real constants that have not been determined, $\theta(t)=(\pi/2, 0, \theta_2, \dots, \theta_M)$, $R(t)=(R_0,0,R_2,\dots,R_M)$ [14]. Hence $\mathcal{F}(\zeta)$ satisfy

$$\mathrm{Re}\left(e^{i(0.5\pi-\theta_\iota)}\log\left(\mathcal{F}\left(\zeta_\iota(t)\right)\right)\right)=-R_\iota,\ t\in I_\iota,\ \iota=0,\dots,M, \tag{4}$$

And $\mathcal{F}(\zeta)$ can be reformulated as [14]:

$$\frac{\mathcal{F}(\zeta)}{K_1(\zeta)}=K_2(\zeta)e^{(\zeta-\alpha)\hat{h}(\zeta)+ih_1}, \tag{5}$$

Where,

$$K_1(\zeta)=\begin{cases}\left(\dfrac{\zeta^2+p^2}{2p\zeta}\right)+c, & \zeta\in\Gamma_1,\\ 1, & \zeta\notin\Gamma_1,\end{cases}$$

$$K_2(\zeta)=\begin{cases}1, & \zeta\in\Gamma_1,\\ \left(\dfrac{\zeta^2+p^2}{2p\zeta}\right)+c, & \zeta\notin\Gamma_1,\end{cases}$$

ρ is a radius of Γ_1 c=1 for Ω_1, c=0 for Ω_2, $\hat{h}(\xi)$ is an analytic function in Ω_1 for c=1 and $\hat{h}(\xi)$ is an analytic function in Ω_2 for c=0. And then define $S(t)$ by,

$$\bar{B}\log\left(\frac{\mathcal{F}(\zeta)}{K_1(\zeta)}\right)=r(t)+iS(t),\ \bar{B}=e^{i(0.5\pi-\theta_\iota)},$$
$$r_\iota=-R_\iota,\ \iota=0,\dots,M, \tag{6}$$

We assume that,

$$F(\zeta)=e^{\bar{B}\log\left(\frac{\mathcal{F}(\zeta)}{K_1(\zeta)}\right)}=e^{r_\iota+iS_\iota(t)},\ \iota=0,\dots,M, \tag{7}$$

which implies that,

$$F'\left(\zeta_\iota(t)\right)\zeta'_\iota(t)=iS'_\iota(t)F\left(\zeta_\iota(t)\right),\ j=0,\dots,M. \tag{8}$$

After some algebraic manipulations, we obtain,

$$\overline{\left(\frac{F'(\zeta)}{F(\zeta)}\right)}=-T(\zeta)^2\left(\frac{F'(\zeta)}{F(\zeta)}\right),\ T(\zeta)=\left(\frac{\zeta'(t)}{\zeta(t)}\right),\zeta\in\Gamma. \tag{9}$$

From the definitions of $F(\zeta)$ and $\mathcal{F}(\zeta)$ we obtain,

$$F(\zeta)=e^{B\log(K_2(\zeta))}e^{\bar{B}\left((\zeta-\alpha)\hat{h}(\zeta)+ih_1\right)}, \tag{10}$$

Let,

$$D(\zeta)=\frac{BF'(\zeta)}{F(\zeta)}-\frac{k'_2(\zeta)}{k_2(\zeta)}\text{, is analytic in }\Omega. \tag{11}$$

$$\frac{K'_2(\zeta)}{K_2(\zeta)}=\begin{cases}0, & Z\in\Gamma_1,\\ \left(\dfrac{\zeta^2-p^2}{\zeta(\zeta^2+2cp\zeta+p^2)}\right), & Z\notin\Gamma_1,\end{cases}$$

Combining (9) and (11), we obtain,

$$D(\zeta)=-\frac{B(\zeta)\overline{T(\zeta)}}{\overline{B(\zeta)}T(\zeta)}\overline{D(\zeta)}-B(\zeta)^2$$
$$\overline{T(\zeta)^2}\,\overline{\left(\frac{k'_2(\zeta)}{k_2(\zeta)}\right)}-\frac{k'_2(\zeta)}{k_2(\zeta)},\ \zeta\in\Gamma \tag{12}$$

Equation (10), yields,

$$\log\left(F(\zeta(t))\right)=\bar{B}\log\left(K_2(\zeta(t))\right)+$$
$$\bar{B}\left[\left(\zeta(t)-\alpha\right)\hat{h}\left(\zeta(t)\right)+ih_1\right] \tag{13}$$

We reach the following from (7):

$$\log\left(F(\zeta(t))\right) = r_\iota + iS_\iota(t), \text{ where } r_\iota = -R_\iota,$$
$$\iota = 0,\ldots,M, \tag{14}$$

then

$$\bar{B}\left(\zeta(t)-\alpha\right)\hat{h}\left((t)\right) = r_j + h_1\cos\theta + i\left(\rho(t)+\hat{\upsilon}(t)-h_1\sin\theta\right) - \bar{B}\log\left(K_2\left(\zeta(t)\right)\right) = h(t) + i\left(\rho(t)+\upsilon(t)\right)+\gamma(t)+i\mu(t), \tag{15}$$

where

$$\gamma(t)+i\mu(t) = -\bar{B}\log\left(K_2\left(\zeta(t)\right)\right), h(t) = r_\iota + h_1\cos\theta \text{ and } \upsilon(t) = \hat{\upsilon}(t) - h_1\sin\theta.$$

By obtaining h_ι, ι=0,…,M, from second equation in Theorem 2; Sangawi [11] we obtain

$$r_\iota = h_\iota - h_1\cos\theta_\iota, \quad \iota = 0,\ldots,M \tag{16}$$

By using Theorem 1 in Sangawi [11], (11), (12) and after some algebraic operations we achieve the following:

$$\frac{F'(\zeta)}{F(\zeta)}T(\zeta) + \frac{1}{\pi}\int_\Gamma \operatorname{Im}\left[\frac{\overline{B(\zeta)}}{\overline{B(w)}}\frac{T(\zeta)}{\zeta-\xi}\right]\frac{F'(\xi)}{F(\xi)}T(\xi)|d\xi| = 2i\operatorname{Im}\left[\frac{\overline{B(\zeta)}T(\zeta)K_2'(\zeta)}{K_2(\zeta)}\right], \zeta\in\Gamma \tag{17}$$

Assume that $\xi=\xi(t)$ and $\xi=\xi(s)$. Then by placing $F'(\zeta)\zeta'(t)/F(\zeta) = iS_\iota'(t)$, $\xi\in\Gamma$ in (17), we realize

$$S_\iota'(t) + \int_I \tilde{N}(s,t)S_\iota'(s)ds = 2\operatorname{Im}\left[\frac{\overline{B(\zeta)}\zeta'(t)K_2'(\zeta(t))}{K_2(\zeta(t))}\right],$$

this can be written in its operator form ($\tilde{N} = -N^*$)

$$\left(I+N^*\right)S_\iota' = 2\operatorname{Im}\left[\frac{\overline{B(\zeta)}\zeta'(t)K_2'(\zeta(t))}{K_2(\zeta(t))}\right] \tag{18}$$

As a result, (18) is not uniquely solvable. To deal with this problem, observe

$$\int_I S_\iota'(t)\,dt = 0, \quad \iota = 1,\ldots,M$$

which implies

$$JS_\iota' = 0 \tag{19}$$

The combination of (18) and (19) gives:

$$\left(I+N^*+J\right)S_\iota' = 2\operatorname{Im}\left[\frac{\overline{B(\zeta)}\zeta'(t)K_2'(\zeta(t))}{K_2(\zeta(t))}\right]. \tag{20}$$

In the light of [14, Theorem 2], (20) is uniquely solvable. $S_\iota'(t)$ gives the value of $S_\iota(t)$, $\iota = 0,\ldots,M$, by using the following equation.

$$S_\iota(t) = \int S_\iota'(t)\,dt + \upsilon_\iota =: \rho_\iota(t) + \upsilon_\iota, \quad t\in I_\iota, \tag{21}$$

where υ_ι is a real constant integration, we see that,

$$\rho_\iota(t) = \int S_\iota'(t)\,dt, \quad t\in I_\iota \tag{22}$$

h_ι is obtained through solving (13) and (11) in Sangawi [11] from which r_ι is provided through (16). Having solved (20) we are granted the value $S_\iota'(t)$. We obtain υ_ι through the equations (37), (38) and (12) in Sangawi [11] from which $\hat{\upsilon}_\iota$ is acquired, after that $\mathcal{F}(\zeta)$ is attained by,

$$\mathcal{F}\left(\zeta_\iota(t)\right) = K_1\left(\zeta_\iota(t)\right)e^{B(r_\iota+i(\rho_\iota(t)+\hat{\upsilon}_\iota(t)))},$$
$$j = 0,\ldots,M. \tag{23}$$

Using the Cauchy integral formula, the interior value of $\mathcal{F}(\zeta)$ is determined.

4. NUMERICAL EXAMPLES

Nyström's method alongside the trapezoidal rule [17] [18] was used to solve (13) in Sangawi [11] and (20). The computational details are almost identical to [19], [20].

Some test regions of connectivity three, four, and seven have been used for numerical experiments [14]. MATLAB R2020a was used to carry out all the computations. In each boundary Γ_j the same number of collocation point has been used. Ω, Ω_1 and Ω_2 are shown in Figures 2-4. Tables 1-3 exhibit our computed values of r_ι,ι=0,…,6 compared to those of Nasser [14].

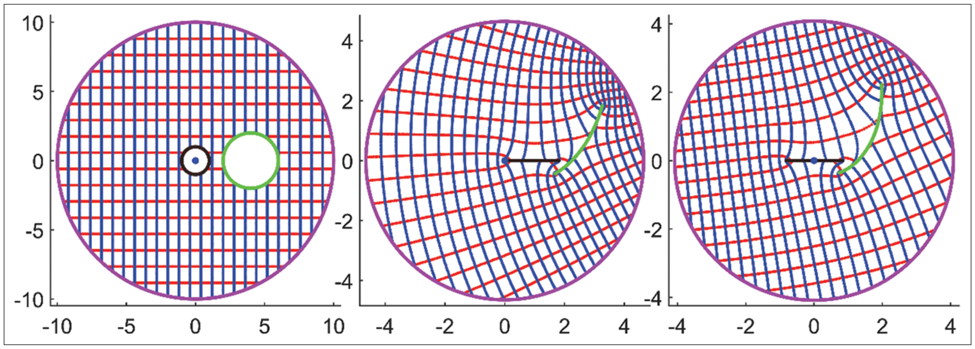


Fig. 2. Mapping Ω onto Ω_1 and Ω_2 with three connectivity.

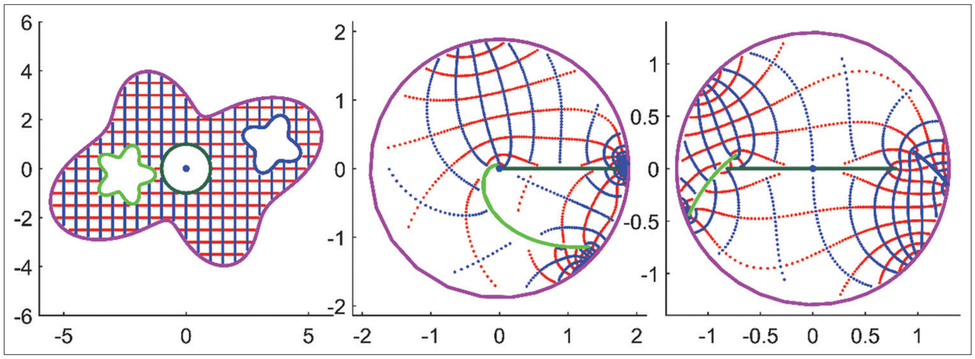


Fig. 3. Mapping Ω onto Ω_1 and Ω_2 with four connectivity.

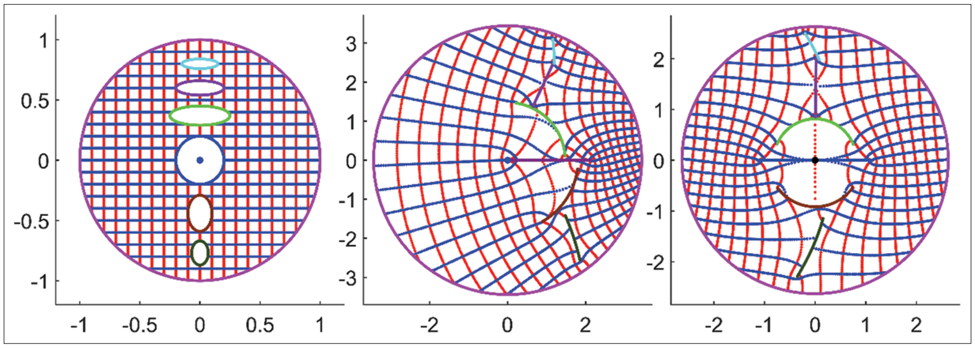


Fig. 4. Mapping Ω onto Ω_1 and Ω_2 with seven connectivity.

Example 1

The region with the following boundaries:

$$\Gamma_0 \;:\; \left\{\zeta(\mathrm{t}) = 10\left(\cos t + i\sin t\right)\right\},$$

$$\Gamma_1 \;:\; \left\{\zeta(\mathrm{t}) = \cos t - i\sin t\right\},$$

$$\Gamma_2 \;:\; \left\{\zeta(\mathrm{t}) = 4 + 2(\cos t - i\sin t)\right\},$$

$$\mathrm{t} \in [0, 2\text{À}], \quad = \left(\frac{\pi}{2}, 0, -\frac{3\pi}{4}\right), \quad \alpha = 2 + 3i$$

Table 1 give the computed values of r and Ω, Ω_1 and Ω_2 are shown in Figure 2.

Example 2

The region Ω bounded by four multiply connected region, Nasser [14],

$$\Gamma_0 \;:\left\{\zeta(t) = (4 + cos\;2t + sin\;4t)\, e^{it}\right\},$$

$$\Gamma_1 \;:\left\{\zeta(t) = e^{-it}\right\},$$

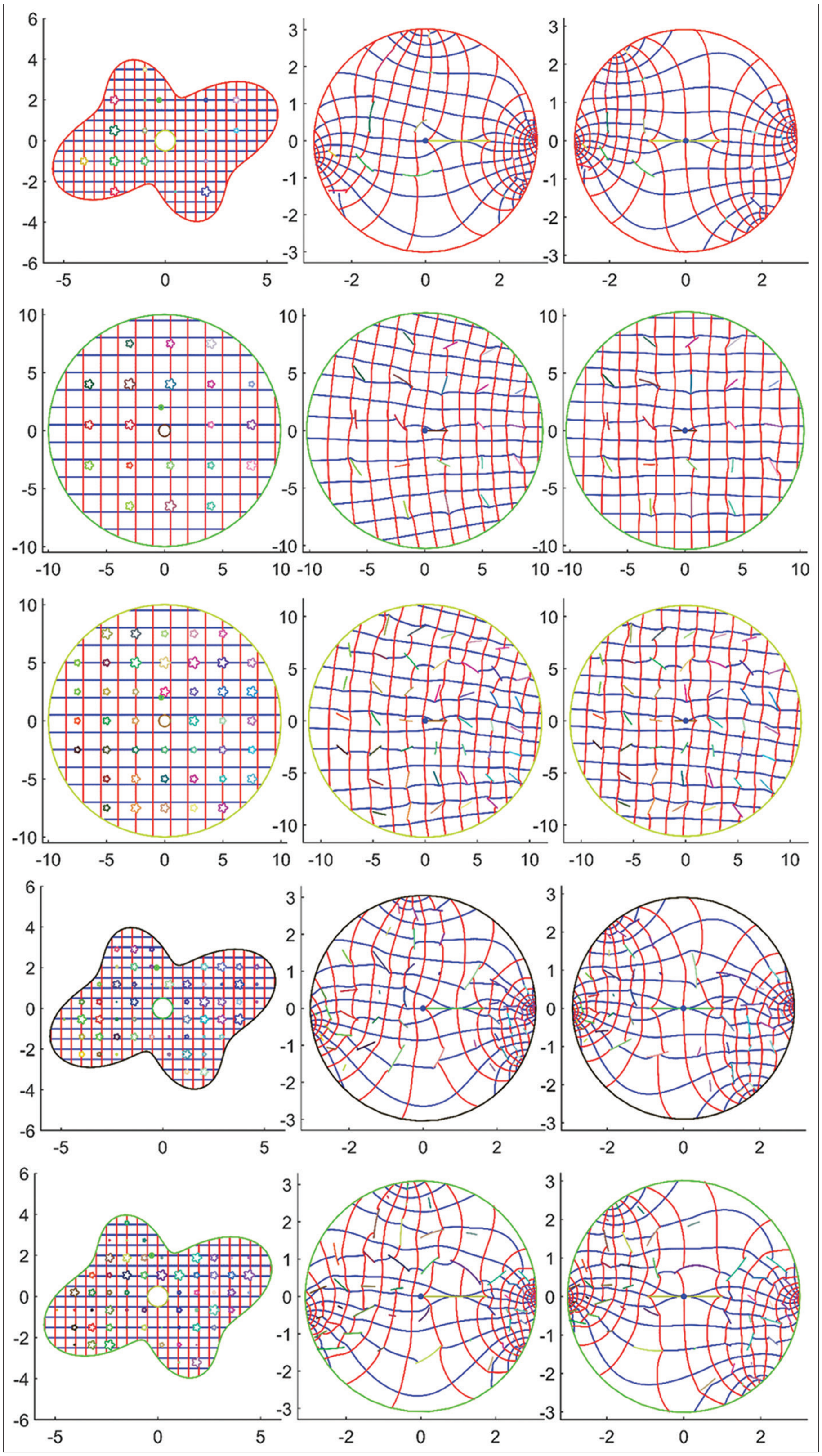


Fig. 5. Ω, Ω_1 and Ω_2 with high connectivity.

Table 1: Computed values of r_ι,ι=0,1,2

Computed values of r_ι,ι=0,1,2 for Ω_1

n	r_0	r_1	r_2
32	1.53579417691881	0	−0.570199621240809
64	1.53579417442883	0	−0.570199620799919
128	1.53579417442883	0	−0.570199620799919

Computed values of r_ι,ι=0,1,2 for Ω_2

n	r_0	r_1	r_2
32	1.40674522978284	0	−0.174539534772558
64	1.40674523035241	0	−0.174539535774301
128	1.40674523035241	0	−0.174539535774301

Table 2: Computed values of r_ι,ι=0,…,3 with n=128

Computed values of r_ι,ι=0,…,3 for Ω_1

r_ι	Proposed method	Presented method in Nasser [11]
r_0	0.632297599993722	0.632297593480685
r_1	0	0
r_2	0.89615354957018	0.896153544382371
r_3	0.47224415769662	0.472244154231282

Computed values of r_ι,ι=0,…,3 for Ω_2

r_1	Proposed method	Presented method in Nasser [11]
r_0	0.260931620532661	0.260931605731473
r_1	0	0
r_2	2.13317552874793	2.13317551821617
r_3	0.100276605317663	0.100276596055608

Table 3: Computed values of r_ι,ι=0,…,6

Computed values of r_ι,ι=0,…,6 for Ω_1

r_ι	Proposed method	Presented method in Nasser [11]
r_0	1.23570971232484	1.235709712326
r_1	0	0
r_2	0.390611187505815	0.390611187504613
r_3	−1.12990942307151	−1.12990942307163
r_4	−0.537364361903202	−0.537364361903388
r_5	0.590670832485013	0.590670832470072
r_6	−0.43049693297431	−0.430496932984193

Computed values of rι,ι=0,…,6 for Ω_2

r_ι	Proposed method	Presented method in Nasser [11]
r_0	0.967817156520659	0.967817156521745
r_1	0	0
r_2	−0.202389826842849	−0.202389826843975
r_3	−1.56125001631144	−1.56125001631162
r_4	−1.05242061198575	−1.05242061198594
r_5	−0.0844703418222131	−0.0844703418361004
r_6	−1.2703736171962	−1.27037361720976

$$\Gamma_2 \ :\left\{\zeta(t) = \left(1 + 0.25\, cos\ 5t\right) e^{-it}\right\},$$

$$\Gamma_3\ :\ \{\zeta(t) = (1 + 0.25\, sin\ 4t)\, e^{-it},$$

$$\mathrm{t} \in [0, 2\pi], = \left(\frac{\pi}{2}, 0, \frac{\pi}{4}, \frac{3\pi}{4}\right),\ \alpha = 2i$$

The values of r_ι,ι=0,…,3 are listed in Table 2 and Ω, Ω_1 and Ω_2 are shown in Figure 3.

Example 3

The region Ω bounded by seven multiply connected region, Nasser [11],

$$\Gamma_0\ :\left\{\zeta(t) = e^{it}\right\},$$

$$\Gamma_1\ :\left\{\zeta(t) = 0.2e^{-it}\right\},$$

$$\Gamma_2\ :\left\{\zeta(t) = i0.37 + 0.25\cos t - i0.08\sin t\right\},$$

$$\Gamma_3\ :\left\{\zeta(t) = i0.6 + 0.2\cos t - i0.06\sin t\right\},$$

$$\Gamma_4\ :\left\{\zeta(t) = i0.8 + 0.15\cos t - i0.04\sin t\right\},$$

$$\Gamma_5\ :\left\{\zeta(t) = -i0.44 + 0.1\cos t - i0.15\sin t\right\},$$

$$\Gamma_6\ :\left\{\zeta(t) = -i0.77 + 0.07\cos t - i0.1\sin t\right\},$$

$$\mathrm{t} \in [0, 2\text{À}],\quad \alpha = 0.5 + 0.5i,$$

$$\theta = \left(0.5\pi, 0, 0.5\pi, 0, 0.15\pi, 0.5\pi, 0.875\pi\right)$$

The values of r_ι,ι=0,…,6 are listed in Table 3 and **Ω, Ω_1 and Ω_2** are shown in Figure 4. Some more examples are shown in Fig. 5.

5. CONCLUSIONS

The present study proposes a new boundary integral equation for the conformal mapping of multiply connected regions onto the disk with rectilinear slit and spiral slits regions, Ω_1 and Ω_2. We used the proposed method to compute several mappings of some test regions and computed the boundary values of the mapping function. The interior mapping function was then determined using Cauchy's integral formula. Numerical examples were provided to demonstrate the high accuracy of the boundary integral equation method.

ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Higher Education through the Charmo University, Kurdistan, Iraq.

This support is gratefully acknowledged. I wish to thank Prof. Dr. Arif Asraf for his cooperation and thank an anonymous referee for valuable comments and suggestions on the manuscript which improve the presentation of the paper.

REFERENCES

[1] P. Koebe. Abhandlungen zur Theorie der konfermen Abbildung. IV. Abbildung mehrfach zusammenhängender schlicter Bereiche auf Schlitzcereiche (in German), *Acta Mathematica,* vol. 41, no. 1916, pp. 305-344.

[2] M. M. S. Nasser and A. A. Al-Shihri Fayzah. “A fast boundary integral equation method for conformal mapping of multiply connected regions”. *SIAM Journal on Scientific Computing,* vol. 35, no. 3, pp. A1736-A1760, 2013.

[3] M. M. S. Nasser. “A boundary integral equation for conformal mapping of bounded multiply connected regions”. *Computational Methods and Function Theory*, vol. 9, no. 1, pp. 127-143, 2009.

[4] M. M. S. Nasser. “Numerical conformal mapping via boundary integral equation with the generalized Neumann kernel”. *SIAM Journal on Scientific Computing*, vol. 31, pp. 1695-1715, 2009.

[5] M. M. S. Nasser. “Numerical conformal mapping of multiply connected regions onto the second, third and fourth categories of Koebe’s canonical slit domains”. *Journal of Mathematical Analysis and Applications*, vol. 382, pp. 47-56, 2011.

[6] M. M. S. Nasser, A. H. M. Murid and A. W. K. Sangawi. “Numerical conformal mapping via a boundary integral equation with the adjoint generalized Neumann kernel”. *TWMS Journal of Pure and Applied Mathematics*, vol. 5, no. 1, pp. 96-117, 2014.

[7] Z. Nehari. “*Conformal Mapping”*. Dover Publication, New York, 1952.

[8] A. W. K. Sangawi, A. H. M. Murid and M. M. S. Nasser. “Linear integral equations for conformal mapping of bounded multiply connected regions onto a disk with circular slits”. *Applied Mathematics and Computation,* vol. 218, no. 5, pp. 2055-2068, 2011.

[9] A. W. K. Sangawi and A. H. M. Murid. “Annulus with spiral slits map and its inverse of bounded multiply connected regions”. *International Journal of Scientific Engineering and Research*, vol. 4, no. 10, pp. 1447-1454, 2013.

[10] A. W. K. Sangawi. “Spiral slits map and its inverse of bounded multiply connected regions”. *Applied Mathematics and Computation,* vol. 228, pp. 520-530, 2014.

[11] A. W. K. Sangawi. “Straight slits map and its inverse of bounded multiply connected regions”. *Advances in Computational Mathematics*, vol. 41, pp. 439-455, 2015.

[12] A. W. K. Sangawi, A. H. M. Murid and L. Khiy. “Fast computing of conformal mapping and its inverse of bounded multiply connected regions onto second, third and fourth categories of Koebe’s canonical slit regions”. *Journal of Scientific Computing,* vol. 68, pp. 1124-1141, 2016.

[13] G. C. Wen. “*Conformal Mapping and Boundary Value Problems.* English Translation of Chinese Edition, 1984. American mathematical Society, Providence, 1992.

[14] M. M. S. Nasser. “Numerical conformal mapping of multiply connected regions onto the fifth category of Koebe’s canonical slit regions”. *Journal of Mathematical Analysis and Applications*, vol. 398, pp. 729-743, 2013.

[15] T. K. DeLillo, T. A. Driscoll, A. R. Elcrat and J. A. Pfaltzgraff. “Radial and circular slit maps of unbounded multiply connected circle domains”. *Proceedings: Mathematical, Physical and Engineering Sciences*, vol. 464, no. 2095, pp. 1719-1737, 2008.

[16] R. Wegmann and M. M. S. Nasser. “The Riemann-Hilbert problem and the generalized Neumann kernel on multiply connected regions”. *Journal of Computational and Applied Mathematics*, vol. 214, pp. 36-57, 2008.

[17] K. E. Atkinson. “*The Numerical Solution of Integral Equations of the Second Kind*”. Cambridge University Press, Cambridge, 1997.

[18] P. J. Davis and P. Rabinowitz. “*Methods of Numerical Integration*”. 2nd ed. Academic Press, Orlando, 1984.

[19] A. H. M. Murid and H. Laey-Nee. “Numerical experiment on conformal mapping of doubly connected regions onto a disk with a slit”. *International Journal of Pure and Applied Mathematics*, vol. 51, no. 4, pp. 589-608, 2009.

[20] A. H. M. Murid and H. Laey-Nee. “Numerical conformal mapping of bounded multiply connected regions by an integral equation method”. *International Journal of Contemporary Mathematical Sciences,* vol. 4, no. 23, pp. 1121-1147, 2009.

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

زانكۆی گەشەپێدانی مرۆیی
جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.7 No.(1) June 2023

2023 2723

e.mail:jst@uhd.edu.iq