Data Mining and Its Effective Role in the Fight Against Diseases and Determine the Causal Relationship

Samar Alsulami
Information Science Dept.– College of Arts and Humanities
King Abdulaziz University - Jeddah - KSA

Abstract. Health data is complex, huge and heterogeneous data, so it is difficult to analyze it by traditional methods, especially as it is data based on tracking and investigation, especially in the field of chronic diseases, where the investigation of the history of the sick person and the history of the social relations associated with him requires investigation.

On the other hand, data mining in the field of non-communicable diseases is the main support for the task of knowledge extraction as this study aims to assess the current state of knowledge in data mining techniques that will help in preventing and diagnosing non-communicable diseases. To achieve the goal of the study, the researcher used a critical evaluation approach. By reviewing the theoretical literature and previous studies related to the subject of the study, it also aimed to identify the objectives pursued by the studies, results, and recommendations, and to indicate aspects of agreement and differences in light of the data.

The current research has reached several results, the most prominent of which is that the health field really needs to search for data due to the tremendous growth of electronic health records. The tools used in data exploration have varied and varied, and the technical systems have also varied. The paper also concluded that the focus of recent studies has been on research in Data mining techniques used to prevent outbreaks of chronic and non-infectious diseases such as heart disease, diabetes and stroke, and that data mining science needs more recent studies because developments in the health field are accelerating and discoveries are successive and on the other hand we notice the spread of some diseases more than before, so the medical field needs For more research on modern techniques to confront and combat infectious and non-communicable diseases as well.

Keywords: Data mining, smart algorithms, communicable diseases, disease control, disease diagnostics, heart disease .

1 INTRODUCTION

The widespread spread of information technology and its ease of availability has led to a huge increase in the volume of data and information that has not been witnessed before in history.
In data, it is derived from many sciences such as statistics, mathematics, logic, learning science, artificial intelligence and expert systems, pattern recognition science, machine science and other sciences that are considered smart and non-traditional sciences. Data mining technology appeared in the late 1980s and proved its existence as one Successful solutions to analyze huge amounts of data, by converting it from mere abstract information to information that can be exploited and used after that.
Health data mining to serve the medical field has become an urgent necessity, as the Social Commission for Asia and the Pacific stated that in one of five Asian countries most lives are lost due to non-communicable diseases such as cardiovascular disease, cancer, diabetes and chronic respiratory diseases, and in Australia the statistics showed Heart and circulatory

diseases are the first major cause of death, causing 33.7% of total deaths. In Africa, statistics also stated that heart disease and circulatory diseases are the third leading cause of death. And the European Public Health Organization stated that heart attacks, strokes and other diseases account for 41% of all deaths.

After reviewing these indicators of deaths, and given that the burden of infectious and non-communicable diseases represents a challenge for researchers, many programs and research are conducted from all over the world on this topic with the aim of predicting them and thus knowing how to combat them using many modern technologies in the field of exploration and It is influenced by many demographic, physiological and behavioral factors. And as it was noted that some studies only talk about one technique used or define one geographical area or one disease state and others, as well as some studies used a number of predictive models to predict only one disease. However, there are cases where an individual suffers from more than one disease, and common risk factors change due to the mutual influence between diseases, and there are challenges that emerge in the methodology of data extraction, the interaction of experimental samples, and issues related to data exploration and its reliability.

Therefore, the researcher decided that the light should be shed in this article about how prospecting tools will have an important and influential role in combating these diseases by focusing on studies that discuss the control of infectious and non-infectious diseases as well as diabetes, heart attacks, cancer, heart disease and epidemics, and how these technologies will be. A role in reducing death rates by using the critical evaluation method through literature review and related studies.

The main question of this work will be what is the role of the data mining technology used in disease control and how can its impact be measured by reviewing the studies and literature that referred to this.

2 PREVIOUS STUDIES

This part aims to discuss and analyze some previous studies related to the topic of the current study and to deal with them with criticism and analysis.

Study (Amini et al., 2013**)** Prediction and Control of Stroke by Data Mining (2013)

This study aims to predict the occurrence of stroke and was conducted in Iran during the period 2010-2011 through which information was collected on 807 healthy and sick people. The study relied on using a standard checklist of 50 stroke risk factors such as history of cardiovascular disease, diabetes, high blood lipids, smoking and alcoholism. The study used data mining tools to analyze data, namely: K nearest neighbor and C4.5 decision tree using (WEKA) The importance of this study is that it is one of the few that has adopted the research in the concept of medical data and health field exploration in the study area, Iran.

The results were as follows: the accuracy of the C4.5 and K nearest neighbor algorithm for predicting stroke was 95.42% and 94.18%, respectively, in groups with risk factors for stroke.

Study (Vijiyarani and Sudha, 2013**)** Disease Prediction in Data Mining Techniques (2014)

In this study, the researcher proposes a model for data mining with the aim of predicting heart disease by identifying the patient's identity through the following model:

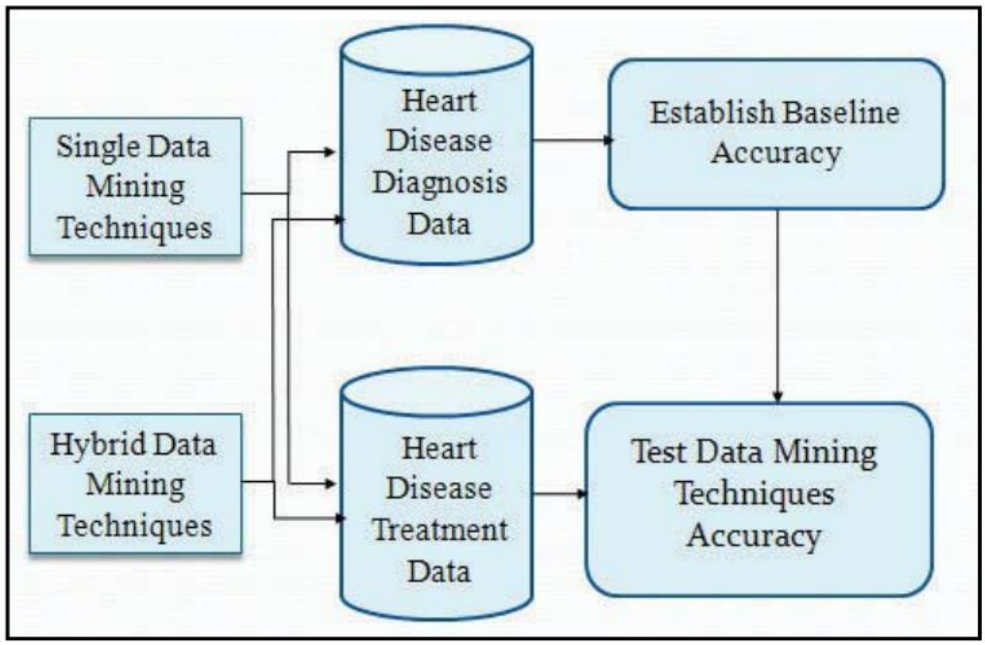

Fig. 1. Study Proposed Model [1]

This model works in 3 main steps:

- Defines criteria for pathogen and baseline data for each
- Application of techniques for exploring individual data used in the diagnosis of heart disease
- Applying mixed data mining techniques to diagnose heart disease
- Apply the same mixed and individual data extraction techniques used in diagnosing heart disease

This study reached the most important outcomes:

- Developing a prototype for a heart disease prediction system using three data mining techniques.
- The system extracts hidden knowledge from the history of heart patients.
- . DMX query language and functions are used to build files and access forms.
- Ease of the form and its answer to complex questions, as the form has easy access to detailed and accurate information

Study (Lashari et al., 2018**)** Application of Data Mining Techniques for Medical Data Classification

A Review (2018)

This study examines the current practices and expectations based on data mining techniques by highlighting some studies and literature, and summarizing the advantages and disadvantages of these technologies through the following table:

Table. 1. Advantages and disadvantages of the algorithms used in data mining [3]

Algorithm	Disadvantages	Advantages
SVM Support Vector Machine	Expensive It takes more time in the training process	Accuracy, ease, suitable for solving many problems, unlike other methods.
Decision Tree	Correct output depends on the correctness of the entered data	There are no requirements before the construction phase

		Reduces ambiguity of complex decisions Ease of dealing with numbers and sequences
ANN Artificial Neural Network	Difficulty of use	It easily dealing independent and variable data
Bayesian Belief Network	It does not give accurate results in some cases where dependency exists between variables.	Ease, speed, accuracy In dealing with big data.
K-NN k- nearest neighbour	It requires a large storage space Noise sensitivity The testing process is slow	Fast and easy to train

We note that this scientific study relatively analyzes current progress in categorizing medical data. And it found that groups with many complex data have more accurate results using mining algorithms and can be verified by experienced professionals working in the field of clinical exploration and research.

Study (Asma, 2011**)** Decision Tree Discovery for the Diagnosis of Type II Diabetes , 2011
This study aimed to demonstrate that extracting knowledge from information stored in medical databases is important for an effective medical diagnosis, as the author used a decision tree algorithm to predict the development of diabetes using the Indian Pima Diabetes Dataset. The applied approach was used on 724 females carrying 6 characteristics (number of pregnancies, blood glucose level, history of blood pressure, body mass index, body mass index, age).
The Weka program was used to generate the decision tree. This research passes through two stages, namely, pre-processing data in the first stage and building a decision tree in the second stage. The accuracy of the decision tree model was 78.177% in predicting diabetes progression for the experimental women. The disadvantage of this study is that the other important factors are not taken into consideration in this research, which are family history, diet and lifestyle.

Study (Patil et al., 2009): Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction 2009.
This study aimed to learn how to use the MAFIA algorithm and used the applied approach in predicting heart patients by finding the repeated and similar elements between a group of people linked by common factors, and the study reached several results, the most important of which is that the effectiveness of the MAFIA algorithm in predicting heart attacks and the defect of this study is its limitations. In a single algorithm, it also did not provide accuracy.

The study (Abdelghani et al, 2006**)**: Predicting Breast Cancer Survivability Using Data Mining Techniques" 2006.
This study aimed to investigate techniques that help in predicting the survival rate of breast cancer patients through the use of three data extraction algorithms which are the Weka toolkit with Naïve Bayes, the neural network, and the C4.5 decision tree algorithms. They used the SEER database that includes information on cancer location, tumor pathology, stages, and cause of death. They used a pre-classification approach with three variables namely survival time recoding (STR), vital state recoding (VSR) and cause of death (COD). The results were quite satisfactory while presenting a comparison of the effectiveness of each proposed network for such problems, and they found that the decision tree algorithm is more accurate than the rest of the algorithms.

The study (Idowu, 2013**)**: Data Mining Techniques for Predicting Immunize-Able Disease: Nigeria as a Case Study2013.
The study used descriptive statistical analysis and is mostly limited to diseases affecting adults. It aimed to provide a model to predict the vaccinable diseases of children aged 0-5 years. The study discussed the difference in disease rates according to their geographical locations, especially in rural areas. Through this model, it also aims to halve the infant mortality rate in these areas by the year 2015. The model has been adapted and deployed for use in six (6) selected local areas within A and Son State in Nigeria. In the proposed model, 3 algorithms were used (ANN, NBC and DT). The study seeks through this model to prove the ability to vaccinate by predicting the incidence of diseases using data mining tools by revealing hidden details in the database.
The study showed that using this model reduces the infant mortality rate from the occurrence of viral diseases such as: yellow fever, measles, polio and hepatitis B; And infectious bacteria
One of the most prominent results of this study was that the NBC algorithm through the model (A | B) = [P (B | A) * P (A)] / P (B) is less complex and faster than the rest of the algorithms. The study also found that using this model of It would enhance the effectiveness of immunization in Nigeria. The results showed that diseases have peak periods that depend on the spread of the epidemic, hence the need to administer the vaccination appropriately in the right places at the right time.

The study (Ahmed, 2017**)**: Analysis of Data Mining Tools for Disease Prediction 2017
The study aimed to make a comparison between data mining tools on the basis of the accuracy of their classification and to know the role of mining databases using algorithms to extract health information for the patient and to clarify the different stages of discovery by exploration as shown in Figure [2] below.
This study talked about data mining techniques and their role in predicting future disease outcomes based on previous data collected from similar diseases, disease diagnosis based on patient data and treatment costs analysis, and among the most prominent of these techniques or tools are Weka, Rapid miner and orange, and are used for analysis. Health data to predict and access health care and reduce costs and time of illness as well .

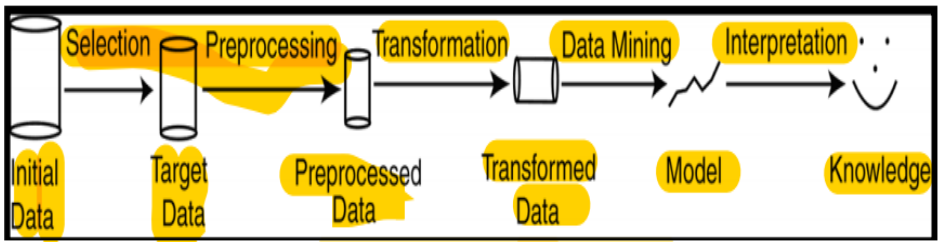

Fig. 2. Patient health data knowledge processes [8]

Below we list the most prominent exploration tools used and define them.

Table. 2. Definition of the most prominent programs used in data mining

Data mining program	Definition
WEKA	Open source software and device from the University of Waikato, WEKA supports many functions such as: data pre-processing, clustering, classification, regression, visualization and feature

	selection. Algorithms can also be implemented using WEKA with data mining and machine learning techniques WEKA provides various sources for downloading data, including files, URLs and databases. It supports file formats including WEKA's ARFF, CSV, Lib SVMs, and C4.5. The WEKA tool includes open source, platform independent, portable GUI and files for a wide range of different data mining algorithms. WEKA is the best tool for beginners as it has many built-in and beta features
RAPIDMINER (RM)	Open source software It provides a good environment for data mining operations. It has a drag and drop feature that is used to build a data flow. Supports various file formats. Data can be imported from various conventional and standard databases
ORANGE	An open source data extraction tool developed In the laboratories of the University of Ljubljana. Applications can be executed using Scripting and visual programming. This tool is suitable for machine learning algorithms and data mining. It can be used easily by researchers in data extraction and Inexperienced users who want to develop and test their own products Algorithms. It gives the advantage of reusing the largest amount of codes.
KNIME	The purpose-built open source data extraction tool developed and maintained by the Swiss company. It is implemented on the Eclipse platform and has facilities for data integration, processing, exploration and analysis platform. KNIME can integrate with other data mining tools such as R and WEKA.

One of the important results that came out of the study is that the tools achieve different results and that WEKA technology is one of the best techniques available for exploration. This is what has been observed by reviewing the studies in this work. Most of the studies indicate the efficiency of this technique. This is because WEKA supports many tasks such as: data pre-processing, clustering, classification, regression and visualization and feature selection.

The study (Chaurasia, 2013**):** Early Prediction of Heart Diseases using Data Mining Techniques, 2017

The aim of this study is to report on the technological developments available to develop predictive models for heart disease survival. Three algorithms, CART, ID3, and DT were used in this study to predict early heart disease. Using three evaluation criteria to find the best algorithm. They found that CART had the most accurate algorithm (83.49%) of the ID3 algorithm as well as the DT algorithm, and the study also resulted in predicting the survival of heart disease as a difficult research problem for many researchers.

3 STUDY METHODOLOY

This study adopted the methodology of critical evaluation for a number of foreign studies, which number (9), and it was obtained through available databases through:

- The Saudi Digital Library

- Google Scholar

Which ranges in the period of time (2009-2018) and it is an approach based on critiquing and analyzing previous studies in terms of the method used, objectives, questions and results, and then critiquing them to reveal strengths and weaknesses and arrive at results that answer the questions of the current study.

4 RESULTS AND DISCUSSION

We find that the study (1), (2), (5) and (9) discussed the outbreak of heart disease, as the four studies shared that they tracked the patient's health history and studied common elements among patients at risk of developing heart disease. These studies were not overlooked in addition to the patient's medical history. To mention the most important techniques used as study (1) used algorithms and an applied aspect, which is the algorithm of the K nearest neighbor and C4.5 decision tree, using (WEKA) program. It compared the algorithms through 3 criteria: accuracy, sensitivity and precision, and the results came as in Figure below.

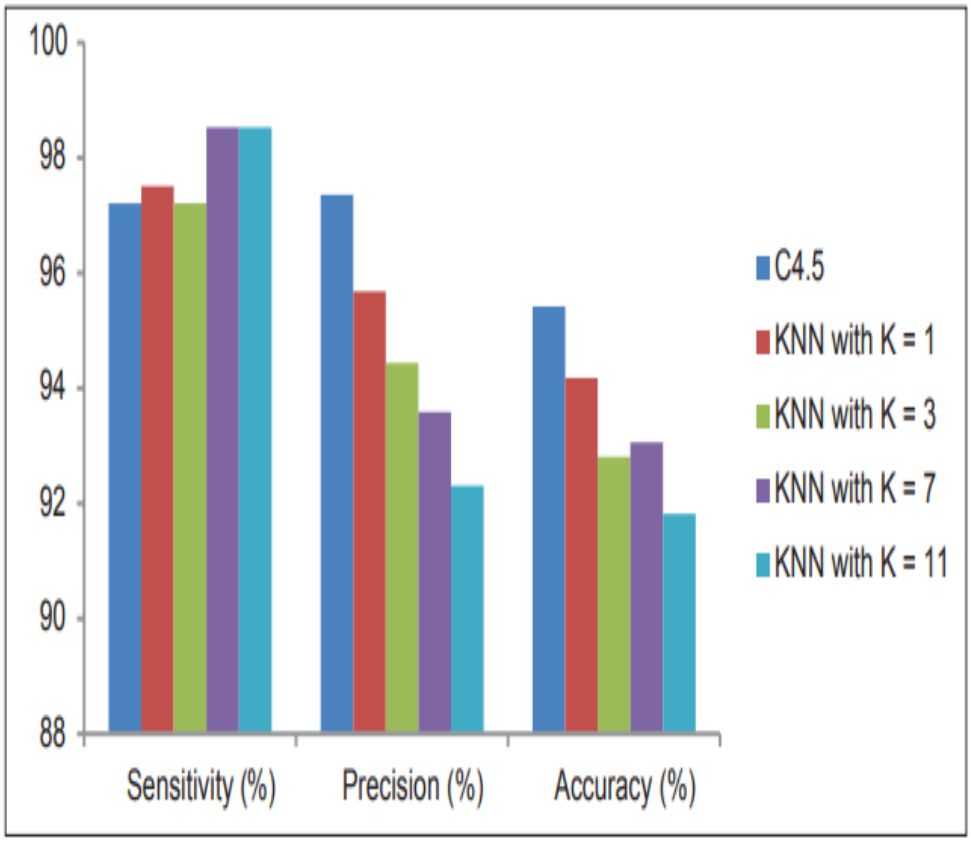

Fig. 3. Comparison of C4.5 and KNN Algorithms [1]

The study (5) used the MAFIA algorithm, and most of this study did not explain how the algorithm works and its effects, as the study showed (1), but it clarified how to predict heart patients by finding repeated and similar elements between a group of people who are linked by common factors as in the following table:

Table. 3. Extracting the recurring factors and their importance

Significant Patterns	Significant Weightage
cholestoral-blood pressure-blood sugar-diagnosis of hd	169.95
thal-sex-blood pressure-blood sugar-diagnosis of hd	167.4
chest pain type-sex-blood pressure-blood sugar-diagnosis of hd	159.53
slope-blood pressure-blood sugar-diagnosis of hd	153.76
exercise induced angina-chest pain type-blood pressure-diagnosis of hd	153.70
thal-chest pain type-blood pressure-diagnosis of hd	152.50
exercise induced angina-blood pressure-blood sugar-diagnosis of hd	151.57
exercise induced angina-chest pain type-blood sugar-diagnosis of hd	142.10
cholestoral-sex-blood sugar-diagnosis of hd	141.51
electrocardiographic-blood pressure-diagnosis of hd	136.4

As for the study (2), it proposed a model for data mining with the aim of predicting heart disease through patient identification and historical medical review.
Study (9), through the use of algorithms, proved that diabetic patients are among the groups at high risk of heart disease, by relying on analyzes of fasting blood and tracking repeated readings by algorithms.

While Study (3) was distinguished from all studies by a frank and direct review of 5 algorithms and mentioned their flaws and advantages, and concluded that the best algorithm is the Decision Tree, and this is confirmed by Study No. (4) where the decision tree algorithm was used to predict the development of diabetes in a group of women. He has 6 characteristics that are expected to be the main cause of diabetes, but this study neglects an important aspect, which is the lifestyle followed by the study sample.

We find that Study (6) is unique in aim from the rest of the studies, as this study aimed to extract data to predict the survival rate of breast cancer patients, but it shares with Study (3) and (4) the algorithm used, which is the decision tree, and it also confirms its preference and accuracy.
The study (7) is unique in its reliance on determining geographical locations, especially rural areas, in the exploration of medical and health data for outbreaks of infectious diseases, and it is the only study that talked about infectious diseases as the rest of the studies were talking about non-communicable diseases. According to its geographical locations
Study (8) shares with Study (1), (3) and (5) in reviewing the techniques used in prospecting, but this study elaborates on explaining the techniques as did the study (3) in listing the algorithms used in punching, as it reviewed the origins and uses of Weka techniques. And Rapid miner and orange .

The following figure shows the classification of accuracy of these mining tools.

We note that studies (1), (4), (6) and (8) referred in one way or another to the use of WEKA technology in medical data mining , and all of them confirmed the effectiveness of this technique as in Figure [4].

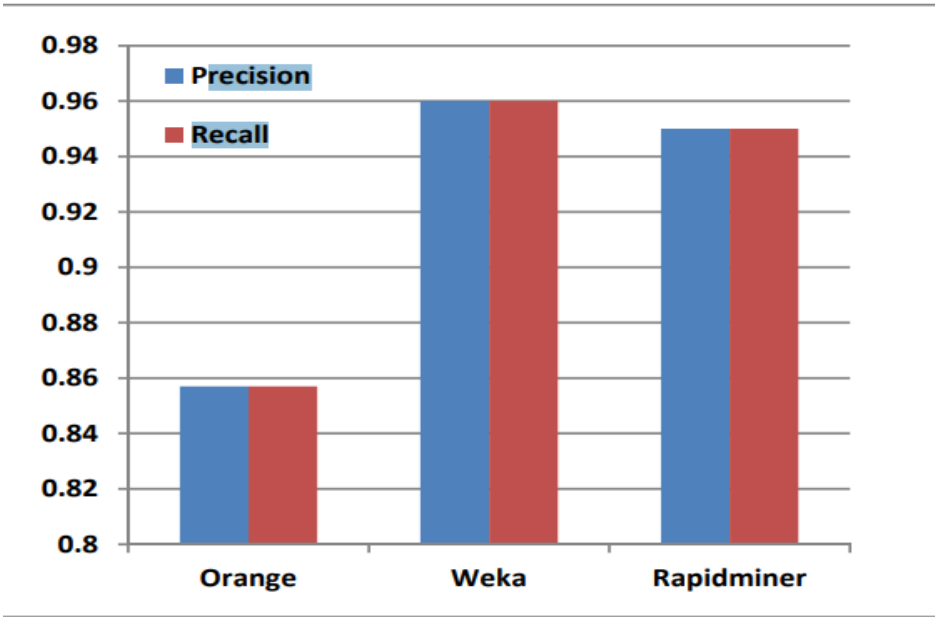

Fig. 4. Accuracy classification of the three mining tools [3]

4 CONCLUSION, RECOMMENDATION AND FUTURE WORK

The health field is one of the most important sectors that really needs data mining due to the tremendous growth of electronic health records, so health care providers and researchers can use data mining from vast stores of data to uncover previously unknown knowledge and then use this information to build predictive models to improve diagnosis and care outcomes. Health.

- The multiplicity and diversity of tools used in data mining, and the multiplicity of technical systems through which these tools operate, as we have explained in this work.
- The focus of the studies is on researching data mining techniques used in preventing outbreaks of chronic and non-communicable diseases such as heart disease, diabetes and stroke.
- Data exploration science needs more recent studies because developments in the health field are accelerating and discoveries are successive. On the other hand, we notice the spread of some diseases more than before. Therefore, the medical field needs more research on modern techniques to confront infectious and non-infectious diseases.
- Some studies have demonstrated the important and effective role of some techniques used in data mining

This study recommends that must follow the following:

- Conducting intensive future studies aimed at preventing the spread of infectious diseases, especially following up on geographical areas that are hotbeds of rapidly spreading viruses.
- After the spread of the Covid 19 virus, we recommend intensifying scientific studies, especially using data mining to combat these rapidly spreading viruses because of their harm to human lives and economic conditions.
- States encourage scientists to research in these areas of research that have a direct impact on health conditions and aimed at improving living conditions.

• Conducting extensive applied research that deals with data mining and its applications in the Arabic language, due to its scarcity.

References

Abdelghani, B. & Erhan, G. (2006). "Predicting Breast Cancer Survivability Using Data Mining Techniques", *Citeseer*, 2006.

Asma A. A.A. (2011). Decision tree discovery for the diagnosis of type II diabetes. *In 2011 International conference on innovations in information technology* (pp. 303-307). IEEE.

Ahmed, K. P. (2017). Analysis of data mining tools for disease prediction. *Journal of Pharmaceutical Sciences and Research*, 9(10), 1886-1888.

Amini, L., Azarpazhouh, R., Farzadfar, M. T., Mousavi, S. A., Jazaieri, F., Khorvash, F., ... & Toghianfar, N. (2013). Prediction and control of stroke by data mining. *International journal of preventive medicine*, 4(Suppl 2), S245

Chaurasia, V. & Pal, S. (2013). "Early Prediction of Heart Diseases Using Data Mining Techniques," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208-217, 2013.

Idowu, A. P., Kayode, A. A., Akhigbe, B. I., Osungbade, A. F., & Adeosun, O. O. (2013). *Data mining techniques for predicting immunize-able diseases*: Nigeria as a case study 1.

Lashari, S. A., Ibrahim, R., Senan, N., & Taujuddin, N. S. A. M. (2018). Application of data mining techniques for medical data classification: a review. In *MATEC Web of Conferences* (Vol. 150, p. 06003). EDP Sciences.

Patil, S. B., & Kumaraswamy, Y. S. (2009). Extraction of significant patterns from heart disease warehouses for heart attack prediction. *IJCSNS*, 9(2), 228-235.

Vijiyarani, S., & Sudha, S. (2013). Disease prediction in data mining technique–a survey. *International Journal of Computer Applications & Information Technology*, 2(1), 17-21.