ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 5, 193-206 (2019)

www.astesj.com

Special Issue on Advancement in Engineering and Computer Science

ASTESJ

ISSN: 2415-6698

Prediction of Non-Communicable Diseases Using Class Comparison Data Mining

Ammar Al-Dallal*,[1], Amina Al-Moosa[2]

[1]*Ahlia University, Computer Engineering Department, P.O. box 10878, Bahrain*

[2] *Ahlia University, Information Technology Department, P.O. box 10878, Bahrain*

ARTICLE INFO

Article history:
Received: 30 April, 2019
Accepted: 07 September, 2019
Online: 08 October, 2019

Keywords:
Communicable disease
Data mining
Diabetes
Predictive data mining
Heart disease
Class comparison

ABSTRACT

Data mining is recognized as an effective technique for extracting and retrieving valuable information or decision from the vast available data. Because of the nature of the functionality of medical centers and hospitals, their data centers contain a collection of valuable information about their patients. By properly processing these data, different applications can be developed to utilize them. These applications could participate in predicting and diagnosing particular diseases. Two prime diseases realized to impact the overall health of society are heart diseases and diabetes. The presented work intends to develop and test a software application that helps doctors and practitioners predict the emergence of noncommunicable diseases (NCDs) such as diabetes and heart diseases. The application applies the predictive data mining model to the medical records which are collected from the Bahrain Defense Force Hospital (BDFH). The BDFH doctors evaluated the application and executed it on actual patients. The results obtained are accurately matching the expectation of doctors in BDFH. All kinds of risks are categorized appropriately according to the defined categories. As a conclusion, this application can help doctors in making proper decisions toward patient health risks. In addition, data mining is more supportive for the health sector and is essential for exploring the knowledge to be used in the health care sector.

1. Introduction

The term data mining refers to the process of knowledge discovery discipline in databases (KDD). It is associated with various computing fields; such as databases, artificial intelligence, as well as software engineering. It can also be used to overcome various problems that emerge where a large volume of data is involved.

The World Health Organization (WHO) is the world body that monitors international health. It showed that 68% of all mortality around the globe is due to noncommunicable diseases (NCDs), and the major killer NCDs are cardiovascular diseases (CVDs), diabetes, lung diseases, and cancer [1-2]. It added that heart diseases cause 12 million deaths globally. CVD comprises of various heart-related disease and its functioning problems that contribute significantly to the adult mortality in some countries.

The (popular) disease or new-generation disease is diabetes. It has emerged as a consequence of inactivity, fewer movements, sedentary lifestyle, and unhealthy eating habits. Given the low activities and other lifestyle reasons, the pancreas produces insufficient insulin to control blood sugar, resulting in diabetes. In a small country in the population such as Bahrain, these two NCDs present a multifaceted problem that must be addressed [3]. Hospitals and health centers accumulate a massive amount of patient's data. This data can be entered or fed to a data mining engine to help in predicting and diagnosing various types of diseases.

The accurate and efficient prediction is a consequence of medical diagnosis. Therefore, it is an important task at that stage. Unfortunately, not all doctors hold expertise for various specialization. In addition, the number of specialized practitioners is also found to be insufficient in many health centers, especially when it is related to heart diseases. Hence, there is a great need to develop a system that assists in predicting and forecasting the patient's diagnosis. It would certainly bring plenty of relief and help doctors.

[1] Corresponding author: Ammar Al-Dallal, Computer Engineering Department, Ahlia University, Manama, Bahrain. Email: aaldallal@ahlia.edu.bh, as_aldallal@yahoo.com

https://dx.doi.org/10.25046/aj040525

This research is designed to assess the data of the patient concerning his health which is necessary to forecast NCDs using inferential data mining and KDD to assists the healthcare professionals in evaluating, as well as predicting, the recurrence or occurrence of NCDs. As a result, this research offers a platform for doing such a system to be able to develop all sets of rules and make them available to medical doctors and other practitioners to anticipate with a degree of correctness the prospect of NCDs among patients. To be able to detect the prime elements which are vital for NCD prediction, the development of an application takes place to compile the massive volumes of patient data for forecasting the potential future trends by employing data mining. Afterward, its testing occurs, which allows the physicians and experts to use the information pertaining to NCDs for improving its diagnosis and reducing the diseases. This project utilizes huge amounts of data obtainable from the BDFH to anticipate NCDs by employing data mining technique.

The contribution that the study adds is that it develops a software which predicts the diagnosis and is tested by BDFH actual practitioners. Using the software, the practitioners were gratified with the study results.

1.1. Purpose of the Study

The purpose of the study is to apply KDD and execute the inferential data mining technique to examine health data, which is essential for predicting NCD. This assists doctors in analyzing or even predicting the recurrence or occurrence of NCDs in patients. Currently, BDFH has no tools or techniques to undertake this task. Hence, a Predicting NCD Application (PNCDA) software will be developed by applying the inferential data mining technique on the compiled huge volumes of available vital data. This application will be available for doctors to be able to predict future trends with accuracy, for the NCD potential among the BDFH patients. Following it, the application will be assessed for demonstrating the model which enables the clinicians and other stakeholders to have access to this facility and use it during their diagnosis process to improve patients' health and to reduce such diseases.

Because of the sensitivity of this project as it deals with confidential patient data, two types of data will be utilized: secondary and primary data. Secondary data will be extracted from the records of the BDFH, while primary data are gathered in person where the related data are organized as per the software requirements. The approach of the interview will be used for collecting primary data, which will recruit about 30 percent to 40 percent of doctors, nurses, as well as BDFH paramedics.

1.2. Study Significance

The objective of this project is to experiment as well as assess different algorithms of data mining, which will assist in NCDs prediction, comparison, and contrast of the effective ways of predicting the disease. The experiment is assumed to serve as an instrumental tool for the doctors for predicting in complex medical cases related to NCDs and advice their parents as per the predictions generated through accurate algorithms, that will have a huge advantage for the field of medical science. This research would provide evidence that the technique of data mining is effective for physicians for predicting and forecasting risky medical cases as practiced across developed nations. This research is set to act as a predecessor for accumulating data for patients in the Kingdom of Bahrain in the future.

1.3 Study Limitations

The study has certain limitations such as it includes patients recorded in BDFH only in Kingdom of Bahrain. Once completed, the findings of the research would be communicated and discussed with the healthcare authorities in the kingdom, where additional comprehensive research can be conducted for covering the whole Bahrain population.

This whole research is categorized into five sections; where section two presents a comprehensive review of the data mining methods and its implementation in the medical discipline. Following it, section 3 provides an explanation of the process for data mining, which is used for the prediction system for NCD. It also provides a description of the prediction software, which is developed. The developed software for prediction and its results are reflected and analyzed in section four. Lastly, section five briefly summarize the overall findings of the research and provides a direction to the future researches for expanding the research horizon.

2. Theoretical Background and Literature Review

In the present times, an increased acceptance of the data mining on the international forums is being recognized, across different medicine and life spheres. Considering the dynamics scope of data mining related to its efficacy for enhancing the healthcare outcomes, this section intends to highlight the theoretical basis which assists in NCD determination as well as the application of the data mining in forecasting.

2.1. Non-communicable Disease

At present, the unbridled growth of NCDs is recognized as the primary cause of mortality across the world. It is reasoned that the increase in sedentary lifestyle and lack of exercise has added to its increase. As it is well-recognized that an active lifestyle is vital for proper maintenance and functioning of the human body. According to 2016 WHO statistics, more than 15 million deaths occurred as a result of NCD such as diabetes and cardiovascular disease [1]. The most surprising and alarming finding of this statistics was that almost half of the deaths were of individuals who were less than 70 years of age.

A. Diabetes

The statistical evaluation of World Health Organization (WHO) findings, and its comparison with the outcomes of 2014, it is found that diabetes had an increasing percentage of about 409 percent for 34 years, ranging from 1980 to 2014. The main contributor to the increasing percentage is the changing world dynamics where the living status of the majority of the countries has changed for both developing and underdeveloped countries [1]

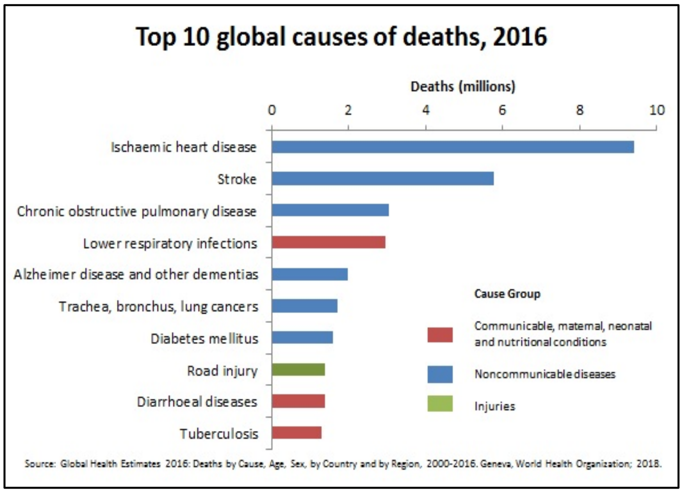
As per the healthcare professionals, the disease of diabetes occurs when the gland of the body (pancreases) is able to release an adequate amount of insulin. It affects the human body function as insulin is responsible for carrying sugar from the bloodstream to numerous cells in order to be used as energy. The deficiency of insulin in the body makes its natural functioning difficult. As a consequence, the body releases high levels of glucose in the urine. Its prevalence for a long-term can cause an organ to fail, cause CVD, and affect the other functions of the body. The deteriorative effects which emerge as a consequence of diabetes are the reason it is ranked as the fourth primary NCD by WHO [2]. It can also lead to the complication of CVDs. The WHO organization [2] documented that diabetes and CVDs were the reasons for more than 11 million fatalities that occurred before the age of 70, as shown in Figure 1.

The primary factor which contributes to the occurrence of diabetes includes the uncontrolled level of blood glucose. Using data mining techniques, the healthcare professionals across the world would be able to forecast illness in an effective manner and will be able to integrate better management technique for patients at high risk. This adds to the significance and needs for the disease analysis and predictions for overcoming it and providing relief to the majority of the patients.

Similar to other regions, Bahrain also lists diabetes as a major health concern. The prevalence of diabetes is found across different ages and populations comprising of different characteristics. The report by WHO for Bahrain documented that diabetes account to 13 percent of the deaths in the region. It also indicates the increasing detrimental outcomes of diabetes, which continue to grow at an increasing speed [3].

Generally, diabetes is categorized into two types; namely, type 1 and type 2. In type 1, the diabetic patient is required to be infused with artificial insulin using medicines and injections. Whereas, in type 2, the body gland (pancreas) produces insulin but is inadequate for the body. Majorly, type 2 diabetes is more prevalent among patients across the world. The population that is generally found to be affected by it comprise of adults, particularly people, in the middle-aged group, however, with the changing lifestyle and dynamics, its prevalence is found in children also.

The low prevalence of Type 1 diabetes is due to its interconnectivity with the external environment, which affects the body insulin-releasing cells. Though, the change in the lifestyle such as regular exercise and maintenance of adequate body weight can help prevent diabetes of type 2.



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

Figure 1: Primary Ten Causes of Death as Reported by WHO [1]

International Diabetes Federation has listed down following practices for preventing against diabetes:

1. It recommends that individuals must start a regime for losing some weight to overcome the effects of diabetes by enhancing insulin resistance and mitigating the prospects of hypertension. So, the people who are overweight are encouraged to sustain adequate body weight.
2. The consistency of physical exercise is integral for sustaining weight loss. It is because the indulgence in the physical activity overcomes arrhythmia, hypertension, and insulin sensitivity; improving the composition of the body; and developing psychological health.
3. Sustaining a healthy diet also overcomes the CVD related risk factors.
4. Smoking, depression, stress, and inadequate sleep can also be considered risky behaviors; therefore, it is preferable to avoid them.

B. Cardiovascular Diseases

Cardiovascular diseases comprise a number of disorders that are related to the heart and blood vessels for example;

Coronary heart disease which is caused by the clogging and narrowing or blockage of the blood vessels which supply the blood to heart muscles.

Cerebrovascular disease is caused by the issues in the arteries, which can affect the supply of blood to the brain. Example of the cerebrovascular disease is a stroke.

The peripheral arterial disease occurs due to plaque (calcium, fats, and other substances) that build up in the arteries supplying the blood to the head, limbs, and the other organs.

Rheumatic heart disease is caused because of the streptococcal bacteria, which attacks the tissues of the body, particularly of the brain and the heart.

Congenital heart disease is caused because of the anomalies of structure and the malformations at the time of birth.

Deep vein thrombosis and pulmonary embolism are caused due to the clotting of blood in the veins of the leg, which might be transferred to the organs like the lungs.

Myocardial infarction (MI) is the term which is used in the medicine for the NCDs in common, which also known as the heart attack. Cerebrovascular accident, which is also known as stroke, is also a kind of NCD and is a serious one. These illnesses or disease occur due to the defects and the faults of the heart and the arteries. The arteries are the vessels for the pumping of the pure blood from the heart into the body. Lack of exercise, bad eating habits, and fat accumulation in the body lead the fat cells to get deposited in the arteries' inner walls. Misuse of alcohol, the use of irregular tobacco, habits of eating, hypertension, and a host of circumstances and conditions lead to the CVA or MI.

Building up of fat inside the blood vessels and the arteries leading to the gradual vessel clogging. Without any treatment, medical care, or the changes in habits of eating or the lifestyles, it can lead to complete blocking of the blood flow. The unmanaged glucose levels in the blood, physical inactivity, and obesity are very common in the population nowadays. Formerly, people had a walking habit, they used to work in the farms, and the field were heavily involved in labor physically, which is rapidly reducing nowadays. Man, in search of comfort and the material gains, has invented a lot of devices and machines, which has reduced the physical activity of humans. The emergence of the fast-food culture is also a danger to health and body, which most of us fail to understand.

If once diabetes, hypertension or the cholesterol imbalance is diagnosed, it is the high time to make sure that the person has the information regarding the upcoming dangers which might lead to the NCDs. These intermediate-risk factors defined by WHO must be highlighted by the primary health center and the clinics.

It is a general knowledge for the people who are having awareness that to stop or to reduce the dependency on the tobacco or the alcohol, to limit the oily and the greasy food that contains high fat and reducing sugar and the salt intake can help them in reducing the chances and the risks which are linked with the NCDs and the CVDs specifically. When a possible danger is detected, the patient should seek medical help immediately and make his lifestyles better. Further, people must use the proper medication for the control or restrict the damage that the factors can cause.

In most of the cases, there are no clear and visible indications for the MI and the stroke as well. It happens all of a sudden, and are not aware of it. Most of the MIs occur when people are sleeping. When the people feel the discomfort, and the pain in the chest or an extreme pain shoots up in the area of the right shoulder or jaws or the elbow, it can be considered as a warning that it can lead to developing of MI or CVA. The only option in this scenario is immediate medical attention.

The common symptoms which are linked with the CVAs are the insensitivity or the numbness in legs, arms, or face as well and it can be on one complete side of the body. There can be disorientation of speech, and even the difficulty in the vision as well. Severe dizziness or hallucinations or the headaches can happen, and in the intense cases, the patients got fainted too.

The third world is mostly affected by the NCDs; it is populated by middle and the low-income groups. The records of WHO speaks for themselves only. There is the accessibility to the medical care for the rich people while the people with the low or middle-income groups, if we look at the profiles of the country, have no or little access to the primary health care. In the countries across Southern America, Asia, and Africa, to approach the medical center is costly, which causes the late detection of NCD for the poor people.

According to the fact sheet of WHO [1], 26% of deaths in Bahrain are caused due to the CVDs. These statistics highlight the possibility of 13% deaths between the ages of 30 to 70 years and is caused by four NCDs. The physicians should record the factors of risks for the CVDs in order to reduce the strokes or the heart attacks by getting the right system for the storage of medical record and the analysis of the data.

C. Diabetes and cardiovascular disease: double jeopardy

Diabetes mainly contributes to the CVDs as proved by the clinical trials and the situation of the people who have suffered from CVDs. Due to the defects of crucial organs like kidneys and the liver, it becomes more difficult to pump the blood for the heart. The defects or the failure of kidney, liver, and the pancreas cause the threatening amount of toxins to sustain in the bloodstream.

The world of medicine has the challenge to handle this double-edged sword. This is the crucial time for the health centers and the governments to look at these two killers; CVDs and diabetes. To aid in reducing the deaths which are caused due to CVDs, data mining methods must be applied to predict the accurate occurrence or the reoccurrence of the CVDs and diabetes.

2.2. Evolution

Organizations nowadays produce a substantial amount of data specific to the institute. For improving the practicability concerning the use of data, researches have introduced algorithms which allow micro-focus on the data and position it as per the super-specific requirements. This advanced the efforts for the development and creation of the machine language algorithms which are useful in the analysis of the different analysis types and formation of decision without or little human supervision. Thus, the evolution of data mining is based on human needs, which assists in the identification of the relationship patterns and forecasting based on the presented layout of the program rules as well as stipulations.

The data mining is described by the researchers and practitioners using various terms. The concept of discovering knowledge from databases has been evolved from data mining. Earlier research of Fayyad et al.[4] has defined data mining as a procedure in which the data sets are implicitly and which reveal previously unidentified but significant information for effective decision-making. This whole procedure is termed as knowledge discovery in database (KDD).

This procedure can be applied in health care to predict the trends of many kinds of diseases and illnesses. Hence, instead of relying on the knowledge and experience, the data mining technique can be used by the doctors, more precisely for KDD to predict trends that would lead to better diagnoses.

KDD increases the efficiency and effectiveness of doctors by allowing them to treat a large number of patients at a given period. Moreover, such a system increases the opportunity for doctors of the same specialization across multiple firms, locations, as well as countries to e-share medical reports for devising best possible diagnoses in a time-effective manner.

The significance and usefulness of the data are evident from different perspectives. The logical alignment of the irrelevant data can be emphasized on the concealed or undiscovered correlations as well as patterns. This can provide valuable data which is critical for examining the individual as well as health being. The main notion is to briefly summarize the voluminous data and conclude its useful findings and information [5].

Data mining is regarded as a statistical interface, which is inclusive of other interference such as statistics, technology database, pattern recognition, data in machine-readable form as well as intelligent expert system [6].

Several definitions are set for data mining, which are raised according to the area of implication.

According to Krishnaiah et al. [7], the facility of data mining enables the use of data for identifying and using the data set trends. The primary findings of this database are to identify the mechanical or automatic patterns, which require less input as well as efforts from the user.

Recently, the spectrum of data mining has enhanced, which is inclusive of artificial intelligence concepts which help in effective and fast-paced management and visualization of the data [8]. The other definition of data mining is provided by Han et al. [9], which states that it helps in the effective extraction of useful data.

Generally, the actual task of data mining is linked with the mechanical analysis of the voluminous amount of data, that is used to attain information which is not yet discovered. It assists in the identification of the data patterns, its categorization using the cluster analysis, odd records identification which require anomalies detection as well as its associated mining rule or dependencies.

Different statisticians' information system communities, as well as data analysts, use the data mining term. The procedure of KDD is recognized as complete, which requires the attainment of the data or discovery of new information. KDD core concept is of data mining; hence, it can be defined as an application which uses specific algorithms required for the analysis and extraction of the data patterns. It is also recognized as the KDD focal hub. It is inclusive of the intellectual approaches required for the data patterns extraction. From a healthcare perspective, the traditional methods are considered to integrate into statistical procedures for the process control comprising of numerous functions pertaining to the fundamental probability distribution which can be executed successfully for controlling the infection rate in hospitals [10].

Previously, in the 1960s, data mining was perceived to be an looked down by mathematicians as well as to statisticians alike. It was regarded as an unhealthy practice which improved its recognition using the term data fishing and data dredging, based on the hypothetical analysis of the data.

The terms database mining emerged in 1980 by HNC, a San Diego–based company, for describing the Database Mining Workstation [11]. With increasing years such as 1990, the term data mining was appeared to have been accepted as a legitimate phrase for describing the harvesting methods employed at the available data for forecasting the happenings in the future. Actually, different words were used for data mining, such as information harvesting, data archeology, knowledge extraction, information discovery, and so on [12].

The term "knowledge discovery in databases or KDD" was first initiated by Gregory Piatetsky-Shapiro and Parker [13]. Though, the relevance of data mining attained relevance as well as acceptance from the communities concerning artificial intelligence and machine learning. Furthermore, this term was also recognized within the business domain for the fourth estate communities. At present, both the terms, such as "data mining" and "knowledge discovery" are used interchangeably. Later in 2011, data mining was defined as data science.

2.3. Knowledge Discovery in Databases (KDD) Process

Knowledge discovery in databases (KDD) serves as a useful process that is used to extract important information from the expanded data. The information is gathered and filtered to extract only the required information. Information that has been collected is employed in the data set, which is further interpreted to attain a useful understanding regarding the given outcomes.

According to Maimon and Rokach [14], the target goals serve as important factors to successfully employ the KDD process. The process begins through fir, considering over the objectives, however, the final product is achieved in the form of newly developed information. This marks an end to the loop. Next involves data mining segments, where final output is achieved in the form of changes involved in the application domain. The results of the process are evaluated by employing fresh data sources while putting an end to the given loop. The finalizing of the process restarts the KDD process. Han et al. [9] elaborated the overall process through the given figure below:

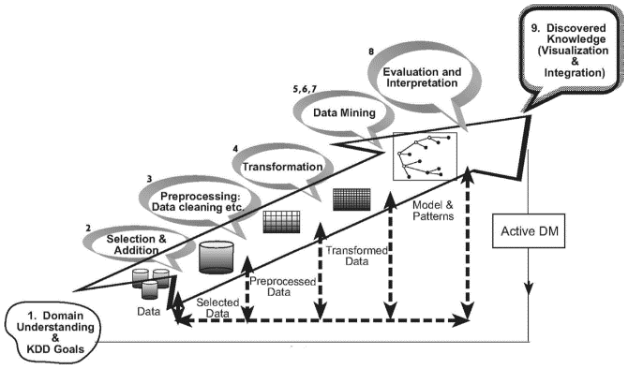

9. Discovered Knowledge (Visualization & Integration)
8
Evaluation and Interpretation
5,6,7
Data Mining
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Model & Patterns
Active DM
Transformed Data
Preprocessed Data
Selected Data
Data
1. Domain Understanding & KDD Goals

Figure 2. KDD process (Maimon & Rokach 2011)

Step 1. Develop a clear understanding of the application domains

The understanding related to the requirements of the healthcare professionals, and end-users, is of foremost value. This helps in providing a clear-cut vision to data miner in meeting the expectations while illustrating important knowledge regarding things that need explanation. It is further important to provide important knowledge regarding things that may help in achieving the desired objective. However, these requirements could be modified after the first round as the end-user might want to add additional functionalities. After identifying important knowledge, it is important for people contributing to the KDD process to function in accordance with the below-given steps. The steps are important for indicating preprocessing measures in the KDD process.

Step 2. Select data sets.

Selection of data sets is important to elaborate on the idea of the target goal. It further helps in providing important details regarding the type and the amount of data that is needed to achieve the target goal. The stage is of greater importance, since any wrong selection of data may lead to several complexities in preparing data. In such cases, the process may result in providing unimportant inferences that may weaken the overall effectiveness of the process.

Step 3. Perform preprocessing and cleansing.

Before processing the selected set of data, it must be filtered and cleaned. The idea is to improve and augment the reliability of the chosen set of data. The factor is important in providing maximum reliability to the selected data set. The cleansing process includes removing noise or barriers that may impede processing; hence, eradicating the given barriers are crucial here. This might serve as a time-consuming process; still, it is effective in indicating maximum surety in terms of validity and reliability of data. In certain cases, insufficient data sets are achieved. However, situations, where ineffective data set, is intervened with accurate predictions, conscious efforts are required to attain expected outcomes. Consider an example; where irrelevant data of patient may act as a barrier in the overall process. Therefore, it is important to remove any such data at this stage

Step 4. Complete data transformation.

The next step is to make the data project-specific. This step ensures the provision of accurate data in a format that produces the required outcomes or results that will be utilized by the doctors to provide accurate predictions. The step is of significant value to provide valuable outcomes in any KDD project. However, important provisions in this regard may help in restoring the KDD process. Two significant methods, including; attribute transformation and dimension reduction are suggested by authors. KDD process serves as a useful tool in providing useful transformations. This helps in indicating maximum validity and efficacy to the formulated results.

Step 5. Choose the adequate data mining task.

Following the above procedure, the next task is to select the data mining type as per the nature of the project. This selection is based on the expected outcomes. Moreover, the selection between regression, and classifications or clustering is based on the desired outcome. Normally, the data mining outcome can be either forecasted, where data mining is done under controls of supervision, or description, where data mining occurs under visualization.

Step 6. Choosing the right algorithm for data mining.

This stage is reflected as the one which discusses the tactics to be used for obtaining the strategic objectives. Is it sufficient to use neural networks? Or is it better to use decisions? In fact, the selection is based on the searching pattern type which is found consistent to the given project and the desired end result. Does it relate to the results precision or decipherability or understandability that forms the project objective? The preference is more towards the neural networks when the precision of the results is required, while the appropriateness of the decision tree is found when gaining an understanding of the patterns and trends is required.

Step 7. Employ the algorithm of data mining.

In this step, the algorithm of data mining is implemented. It could be applied several times for obtaining the best result. For instance, the algorithm could be iterated using different controlling parameters, i.e., the low number of occurrences in a certain leaf or probably executed until the desired accuracy is obtained.

Step 8. Evaluate the mined data.

The mined data evaluation is the core of this KDD stage. This is primarily related to the reasonable interpretation and findings with respect to the discussed and defined goals of the project. This evaluation may promote the need to either add or remove a certain feature from the transformational stage of the data. This stage helps in achieving the usefulness of the data and its comprehension by the end user. Upon finalization, the found information is documented. When the found knowledge is being finalized by the KDD team, the whole data mining procedure is documented and assessed against the predetermined outcomes.

Step 9. Use of the discovered knowledge.

At this stage, the developed KDD process is evaluated for its effectiveness and efficiency by the end user, where refinement occurs in case the fine tuning of the data is required. Also, the created knowledge is being assessed for pilot-testing by practitioners and doctors for ascertaining the production of the desired outcomes. When a new product is being tested in a real environment, some conditions might variate from that at the laboratory. Therefore, these must comprise of the built-in ability to change, adapt, modify the derived knowledge in order to satisfy the end user.

2.4. Data Mining Techniques

Currently, an increased inclination of the researchers is concentrated on data mining. Generally, there are two types of data mining; namely, predictive model and the descriptive model. The two models are as follows;

A. Predictive Model

Prediction, as the name indicates constitutes of the correct envision of the future trend be logical computation of the data. The predictive model uses the previously available information for predicting future outcomes. This model is employed by various firms such as organizations who attempt to data mine the worth of an individual.

The predictive model data mining applies several techniques encompassing regression, classification, time series analysis, and prediction. This model is used for identifying the model that effectively matches the identified ideas or data sets. It is also helpful for class prediction of the objects when there is no availability of the class objects. The achieved model is primarily focused on the assessment of the identification classes set. To assess the numerical values and forecast, the statistical model of regression is used.

B. Descriptive model

This data mining model is employed to identify the data patterns to understand the relationship between the data attributes. The fundamental feature of the data is represented and summarized using the descriptive model. For instance, the customer database and identification of different sets can be useful to the marketers. The identified method can be employed for devising effective marketing programs for targeting audience.

The descriptive model applies several techniques such as clustering, summarization, association rules, and sequence discovery. In the clustering method, the analysis of the data object occurs without reference to the detected class label. Whereas, in the summarization, different properties and specialties of data values are to be considered as noise or outliers.

For the association rules, the same recreated rules are used for assessing the patterns and association among the different obtained characteristics. These are used for analyzing the data patterns for determining or forecasting their classification. Other systematic order identification is through sequence discovery which occurs for identifying the events in a timely manner rules and regulations. Figure 3 demonstrates the classifications of data mining.

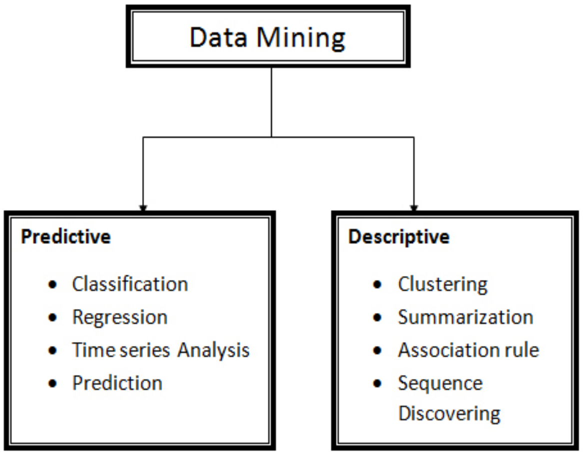


Figure 3. Tasks of Data mining

Among all the models presented in figure 3, the focus of the present study is on one named prediction-mining method.

2.5. Medical Applications Data Mining, Forecasting

Data mining has been widely used in medical applications to solve varies aspects. In this section we highlight some of them.

A. Health care informatics

Medical informatics is termed as dynamic because of the increasingly developing and shifting nature of both medical science and technology. According to Hersh and Hoyt [15], the discipline of health care informatics is related to the resources, materials, tools, and formal ways that are utilized for optimizing the storage, its management and retrieval of the needed biomedical information for effective decision-making, and problem-solving. Likewise, Athina and Konstantinos [16] highlighted that health care informatics employ modern communication, computer applications, as well as IT and computer system in medicine fields in the form of care provided, the delivered education, and research undertaken. Bronzino [17] stated that health care informatics uses different type of tool for using diverse which assists in utilizing and sharing information for improving the delivery of health care. Another term used is medical informatics whereas, it is also sometime referred by practitioners as clinical informatics or bioinformatics. Though, bioinformatics regard it as biological information analysis of the databases that are based on computer and statistical analysis of biological information in a speedy and improved manner. Clinical informatics concerns with maintaining data structure, data organization, and data search to make decisions using relevant data to conduct significant research in the medical discipline.

B. Data mining in medical discipline

Over the years, there is an increase in the computation analysis for deriving computer-based analysis and interference for improving the medical outcomes. The use of software and highly complex computation tools have increased. Previously, healthcare professionals majorly relied on their knowledge, skills, intuition and experience. However, at present, various channels have been developed across firms, regions, as well as states using both formal and informal sharing means for clinical experience which can be employed by the rest of the world. Data mining has provided great benefit to the medical field such as to forecast NCD and major other diseases.

This proliferation of the data mining has increased its efficiency in the medical discipline, where the use of data mining has become a norm where healthcare firms and facilities are being encouraged for meeting the need to predict the trends in disease and their occurrence. The significance of data mining, particularly for information management, is well established. Acharya and Yu [18] the developers of various computer-aided algorithms, have made the health facility more effective and have tremendously improved health care outcomes. Such as, it has aided in informatics, monitoring systems, as well as epidemiology. The difference between the pathological and normal data has facilitated improvement in the diagnosis and decision making.

Recent data mining techniques, as well as the essential statistical tools, can be used to assist doctors in diagnosing diabetes and CVDs. With the use of the statistical analysis, various CVDs encompassing stroke (cerebrovascular disease), cardiac arrests (coronary heart disease), congenital heart diseases, hypertension, peripheral artery disorder, arrhythmia, and heart failure with diabetes can be predicted. An overview of such systems is listed below.

Khaing [19] proposed an efficient approach for assessing and predicting the risk of heart disease. The primary stage is the execution of the k-means clustering algorithm with K=2 for extracting relevant data through clustering the heart disease database. By using k-parameter the number of fragments is mastered. Consequently, the extracted data of heart disease mine the frequent patterns with the use of maximal frequent itemset algorithm (MAFIA). Afterward, the execution of ID3 training algorithm occurs to produce the heart attack level following the decision tree. Although the accuracy of the findings obtained was 74% using k-mean–based MAFIA, there is a need to further increase this accuracy.

Many real-world medical data sets suffer from overlapping information. Overlapping k-means (OKM) is extended from the conventionally used k-means algorithm, and it is recognized as one of the most real-world medical data sets that inherent overlapping information. This clustering method enables one sample to be related to one or more than one cluster. Though, the issue of sensitivity in OKM is prevalent to the initial cluster centroids. Khanmohammadi et al. [20] proposed a hybrid method which assimilates the k-harmonic means and overlapping k-means algorithms (KHM-OKM). In this approach, the initialization of the cluster centers of OKM are obtained from the output of KHM. The performance of KHM-OKM is evaluated in terms of completeness, homogeneity, and cluster size-quantity tradeoff, and it is found that it outperforms the OKM algorithm.

Feature selection is one of the integral steps in data mining. When selected properly, it improves prediction accuracy. Orthogonal Feature Extraction (OFE) is proposed by Jiang et al. [21], which uses the feature ranking techniques. The technique proposed is applied to improve cancer prediction accuracy. To make sure that the attributes of the selected features are accurate, they developed an algorithm that selects the linearly independent vectors iteratively that belong to top-ranked attributes. The results are compared with the analysis of principal component, neighborhood component, and linear discriminant. These three methods are outperformed by OFE in terms of computational complexity and performance.

Another heart disease prediction system is developed by Suvarna et al. [22]. This system combines optimization techniques and data mining. The optimization technique used is particle swarm optimization, which is modified by applying a constriction factor. In this inherently distributed algorithm, the solution is obtained through the interaction among several simple individuals called particles. The authors used *de facto* standard data sets for the reliability ranking of heart disease prediction. The three techniques used for comparing the result are: ID3, multilayer perceptron with backpropagation training and classification, and regression trees. The proposed approach outperforms these three techniques, producing accuracy of classification equal to 53.1%. But the accuracy is much lower than what is expected from actual practitioners.

Rairikar et al. [23] applied the backpropagation technique in genetic algorithm to develop a system that predicts. This prediction system uses 13 attributes obtained from cardiovascular disease information. The results are compared in terms of time complexity with two well-known classification of data mining techniques, which are KNN and Naive Bayes. The results showed that KNN is much better in its performance. However, this method is lagging measuring the accuracy of heart disease prediction.

Tomczak and Zieba [24] applied classification data mining in machine learning to solve two issues that appear when applying machine learning to medical diagnosis. One issue deal with the implementation of interpretable models such as decision tree and classification rules. The other issue deals with the imbalance between classes with high number of examples such as healthy patients and a class with low number of examples such as ill patients. The proposed model is a probabilistic combination of *soft rules*. These rules are constructed by introducing a new random variable called *conjunctive feature*. In addition, a new estimator is introduced to incorporate the knowledge about imbalanced data. This approach is tested using oncology data set. The results show that soft rules can perform well on data sets with small number of examples. In addition, the outcome is comparable with expert oncologists. Meanwhile, this work needs to be examined against large data sets.

Zhu and Fang [25] noticed that the current existing classification trees might not be able to properly provide classification for the provided predictor variables set, which may be categorized using a high error rate. To solve this, they proposed an algorithm that combines the logistic regression model and the trichotomous classification tree. This tree is used to split and establish the tree recursively until it meets the stopping rule of the tree splitting. The algorithm is tested on two real data sets: Pima Indian data set and Wisconsin Breast Cancer data set. The findings showed that the proposed algorithm is more accurate in predicting some diseases than the classification and regression tree. However, this technique suffers several limitations, including using binary response variable, the cost of variables used in classifications ignores the actual cost and thirdly, the variance-covariance between 2 categories use limited simulated normal distributions.

Several algorithms and methods such as regression, clustering, classification, neural networks, artificial intelligence, genetic algorithm, association rules, nearest neighbor, and decision tree have been executed by researchers to assist healthcare practitioners in the effective diagnosis of heart diseases or diabetes to prevent further diseases.

The above-mentioned research studies provide evidence for the effectiveness of the data mining concerning the field of NCDs prediction and correction. The objective of this research is to devise a practical model which can be executed by the healthcare practitioners in the main hospital of Bahrain, or where individuals belonging to particular demographics which adds to its generalizability across the worldwide health care centers.

What features the developed PNCDA is that it allows realistic and timely mean to forecast the general NCD forms, particularly CVDs and diabetes. Rather than using pre-defined datasets, this approach uses real data obtained form BDFH. Further, the developed PNCDA has been tested by actual practitioners from the mentioned hospital and the results met their expectation obtained from the conducted questioner as will be explained later.

3. Research Methodology

As the major outcome of this research is the prediction software system that aids the doctors and practitioners of the BDFH to effectively predict the NCDs of their patients, the first stage of this project is to identify the most effective factors and classify them. These factors must provide good input data (parameter) using an adequate model for data mining. These factors can be identified by the users of the proposed system. Three main factors are included in this research, such as, prediction methods for data mining, patient database, and software application implementation, which are used to characterize and devise a model to predict the cardiac arrest diagnosis as well as diabetes risk among the patients that are listed in the obtained BDFH data set. This research will use both forms of data (primary and secondary). It will integrate in data mining, its significance in healthcare, and medical field application for developing a software system that will be assessed in real situations for evaluating its performance.

3.1 Data Types

In this research, two types of data are used; secondary data and primary data.

A. Secondary data

Secondary data refers to the data that has already been assessed, used and collected by certain firms. In this research, the record of 23000 patients is obtained from BDFH, which are used for designing the required NCD prediction application.

B. Primary data

The data that is collected for the first time as per the defined objectives of the research are termed as primary data. The primary data for the study is collected using a questionnaire which is distributed across 30 specialized healthcare professionals in the hospital. Since all the questionnaire were completed and received; therefore, the response rate is termed as 100 percent. The objective of the questionnaire is to gather the necessary requirement, NCD diagnosis for medical characteristics, which include CVD and diabetes, which are the commonly found forms of NCD.

3.2 Data Collection Procedure

A survey was conducted using the questionnaire as a primary form of data collection. While, for the secondary form of data collection, the health center record was used following the development of the software application.

A. Primary data collection

The questionnaire-based survey was conducted among the doctors to clearly understand the expectations they hold as well as ideas about the current situation concerning NCDs from one side and to examine the awareness and the level of utilization of having a new software application that will help them in predicting NCDs from the other side. To evaluate the current situation concerning NCDs prediction, the BDFH doctors who are either endocrinologists or cardiologists were requested to fill in a questionnaire survey. It was done to gain an understanding pertaining to their perceptions and opinions about the expected software application.

B. Study Population and Sampling

In BDFH, 13 cardiologists and 17 endocrinologists are present. The sample of the study is the entire population given the relative ease to work around. Therefore, the sample would constitute of all the related doctors. The structured questionnaires were administered by the researcher to assess their perception and opinion.

4. Survey Data Analysis and Findings

This section provides information concerning the questionnaire model, which was used in the survey. The survey aimed to establish the importance of predicting model concerning NCD trends among the end-user, and doctors. The direct interviews were held with the doctors to explain to them the significance of developing the software.

The conducted questionnaire was aimed to examine the present scenario concerning the use of NCDs forecasting tools in the BDFH. The sample is formed by the total endocrinologists and cardiologists' populations who are presently on the rolls of the BDFH. As the interviews are conducted on a one-to-one basis; therefore, no interview was missed. The following results were achieved by the questionnaire;

The answers to the first question show that 70% of patients treated by these 30 doctors suffer CVDs and diabetes. This provides evidence that NCDs, more particularly, diabetes and cardiovascular disease are majorly found. The generic and common disease such as influenza cold, allergies, and sprain make a total of about 30 percent. The BDFH does not provide treatment for critical diseases such as cancer.

The second question examines the primary NCDs causes, where the doctors mentioned unhealthy eating habits, smoking, no exercise, and consumption of high-fat food as the major NCDs reasons.

Following third and fourth questions in the questionnaire are related to each other and are also analyzed together. Though 80 percent of the doctors are aware of the prevalence of some form of NCD prediction mechanism across the globe, they firmly say (i.e., 100%) that BDFH lacks any such system, mechanism or facility.

In the fifth and sixth questions, 24 among the 30 doctors (i.e., 80%) showed agreement that presence of computer-based application would assist in predicting such NCDs, which aids the BDFH in notifying patients on a continuous basis. In addition, 90 percent of them welcomed the idea of offering an application that can "correctly predict" the NCDs occurrence, which is essential.

5. Proposed Application System

The software system developed comprises of two main components; the prediction engine and the software application. The system functions as an intermediator (practitioner) between the prediction system and the user. The explanation of each component is presented in which first describes the data that is applied in the proposed prediction system.

5.1 Data Mining Engine

The application purpose is to assess the data of the patients to predict if the patient is at an NCD risk or not. The following tools are used for application programming;

1. VS.Net: Tool for writing the application.
2. VB.Net: Tool for coding the application.
3. Oracle DB: The database which is used for data storage and manipulation.
4. Toad Oracle: Tool for gathering the code and retrieving the data from the database.

Among various techniques for data mining, the research adopted the mining class comparison. This is used to mine a description which compares or distinguishes one class from another related class. It is used for meeting the project purpose based on stages which are followed for the process of mining comparison [9]. The stages are as follows;

1. *Data collection*. It comprises a set of relevant and useful data that is collected by the use of query processing. The segregation of the data occurs in the target class along with contracted classes set.

2. *Dimension relevance analysis*. This is executed only where there are several dimensions and only a few relevant dimensions are to be selected for analysis.
3. *Synchronous generalization*. This is applicable on the target class concerning the degree of control exercised by the user. The contrasting classes concept is to generalize the same level as those in the main target relation, to form the prime contrasting classes' relation.
4. *Presentation of the derived comparison*. The achieved class comparison can be represented in the form of tables, graph, and rules.

5.2 Patient Data Set

The researches, which tackle data mining for medical application use the available data sets such as heart-Statlog Medical data set [26], liver disorder, Indian liver patients, heart disease (Statlog), heart disease (original), hepatitis Parkinson's, Parkinson's, breast cancer Wisconsin (diagnostic), dermatology, and lung cancer [20].

This work is featured using the actual data set obtained from the BDFH, which contains the details and vitals of 23,000 patients.

Using the developed data set, that was extracted with the use of the discussed algorithm of data mining, the recognized pattern assists to predict the risk of getting heart disease and diabetes.

The data of all the patients were processed through the elimination of the repetitive records and completion of the missing details. The dummy values were used for filling the personal details (name and ID number, etc.). Moreover, the recurring patterns are mined with the use of proposed process of data mining. The attributes of the patients gathered from the database are; sex, age, chest pain (CP), FBS (fasting blood sugar), test BPS, rest ECG (electronic cardiographic), smoking, diet, as well as alcohol. Following it, these are used for classifying the patients, which are modified from the study conducted by Khaing [19] as some of them are unavailable in the data set, we have. Afterward, they are adjusted as per the survey conducted with the actual practitioners in the BDFH.

5.3 Application Process Description

The overall exercise was conducted with doctors who will be the main user of the developed software. This is to ascertain that all the unnecessary data is removed so that the prediction of NCDs is barrier-free. The data of the patients is derived from the three tables by executing the second step of KDD. Figure 4 presents the diagram for the entity relationship (ER). The information record of the patients is presented in the first table, second patients' vitals, such as, diet, blood pressure, chest pain type, smoking, and heart rate. The blood test collected from the lab is collected in table 3, where cholesterol and fasting blood sugar are its attributes.

Reflecting that the target class are the lab results and vitals, the "if condition" is used for the constructing class, along with attributes such as sex, age, chest pain (CP), FBS (fasting blood sugar), test BPS, rest ECG (electronic cardio graphic), smoking, diet, as well as alcohol (Table 1). The data mining query language (DMQL) can be used for the data mining task, as follows;

1. Used Patient_DB
2. Mine comparison as "Patient_NCD_Risk_Level."
3. In relevant to sex, age, CP, chol, FBS, test BPS, rest ECG, smoking, diet, alcohol
4. For "Vitals_View" and "Lab_Results"
5. Versus "if_condition_rule"
6. Evaluate sum
7. Show as "risk level."

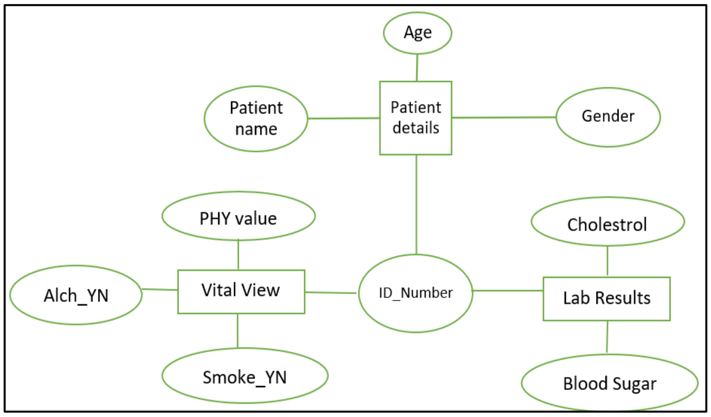


Figure 4. ER diagram of the proposed application

Initially, the change in the query occurs in two forms such as in the two interrelated queries where the two sets of relevant data tasks are accumulating; such as, one is obtained from Vital View table target class and Lab Results table, whereas the other is derived from the contrasting class, i.e., "if condition rule" which is written in the code classes.

Secondly, the two tables, i.e., Vitals View and Lab Results, are analyzed for their dimension relevancy. Furthermore, the weakly relevant dimension are removed, i.e., the removal of the patient information from the found test classes.

Thirdly, the synchronous generalization is applied from the target class to the controlling class level in order to produce the prime relation for the target class. With the use of "if condition" rule, the categorization of the attributes in the form of group is held. Initially, the ages are evaluated. In case, the age is less than 40, it is grouped as young age, whereas, if it lied between 40 to 60, it is grouped as medium age while for 60 or greater than 60, it is regarded as older age. Following it, smokers or alcohol drinkers are categorized. In case the patient is a smoker, he is characterized as group 1, and in group 2 if the situation is vice versa. Similarly, if the patient is an alcohol drinker then he is characterized in group 1, and in group 2 if he does not.

If the blood pressure (BP) of the patient is less than 80, then he is grouped into group 0, and group 1 if it is equal to or more than 80 and group 2 when the blood pressure is higher than 90. Likewise, for the classification of heart rate, if the heart rate is below or equal to 100, it is categorized in group 0 while if it is more than 100 and less then 150, it is categorized as 1. Similarly, in case the heart rate is greater or equal to 150, the patients are categorized in group 2.

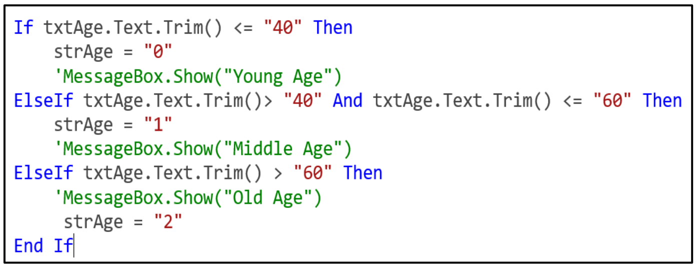
Table 1. Attributes and values used in the proposed application

No.	Attributes	Description	Cut-off Value	Type
1	Age	Age of the	Age <=40 Age <=60 and >40	Young age Middle age Old age

		patients in years	Age >60	
2	Gender	Gender	1 0	Male Female
3	CP	Type of Chest Pain	Value 1 Value 2 Value 3 Value 4	Stable Angina Unstable Angina Nonangina Pain Asymptomatic
4	Test BPS	Resting blood pressure (in mmHg)	BP <80 BP <90 BP >90	Normal Normal to high High
5	Chol	Serum Cholesterol (mmg/dl)	Chol <5.2 Chol <5.2 and >6.2 Chol >=6.2	Normal High Severe
6	FBS	Fasting blood sugar	1 0	True False

7	Rest ECG	Resting electro cardiographic results	Val=0
			Val=1
			Val-2
				Normal
				Abnormal
				Probable

8	Diet	On a healthy diet	0 1	True False
9	Smoking	Smoker patient	0 1	True False
10	Alcohol	Alcohol drinker	0 1	True False

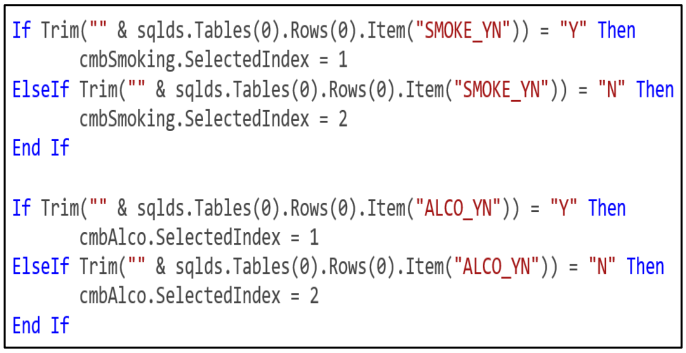


```
If txtAge.Text.Trim() <= "40" Then
    strAge = "0"
    'MessageBox.Show("Young Age")
ElseIf txtAge.Text.Trim()> "40" And txtAge.Text.Trim() <= "60" Then
    strAge = "1"
    'MessageBox.Show("Middle Age")
ElseIf txtAge.Text.Trim() > "60" Then
    'MessageBox.Show("Old Age")
     strAge = "2"
End If
```

Figure 5. Comparison of condition for the age

Lastly, Table 1 presents a comparison of the resulting classes in the rules form. Such as, in case the classes are equal to 0 when there is no risk for patients. While patients are at high risk when the classes are equal to 2. Likewise, the patient is at lower risk, when the classes are equal to one, where the other classes combination occurs based on the recommendation of the actual practitioners.

Following the preprocessing, the previously mentioned attributes in Table 1 are utilized for the code application for using the IF condition. The classification of these values done in the form of cut-off value, following its comparison.



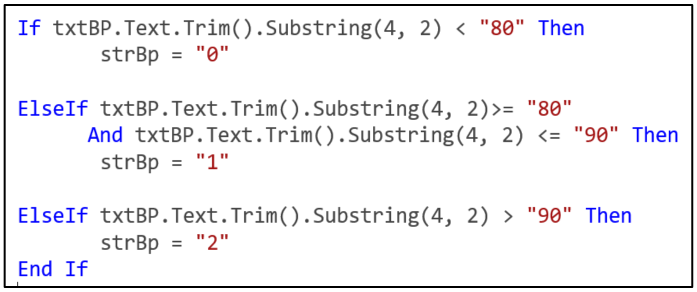
```
If Trim("" & sqlds.Tables(0).Rows(0).Item("SMOKE_YN")) = "Y" Then
       cmbSmoking.SelectedIndex = 1
ElseIf Trim("" & sqlds.Tables(0).Rows(0).Item("SMOKE_YN")) = "N" Then
       cmbSmoking.SelectedIndex = 2
End If

If Trim("" & sqlds.Tables(0).Rows(0).Item("ALCO_YN")) = "Y" Then
       cmbAlco.SelectedIndex = 1
ElseIf Trim("" & sqlds.Tables(0).Rows(0).Item("ALCO_YN")) = "N" Then
       cmbAlco.SelectedIndex = 2
End If
```

Figure 6. Grouping smoke and alcohol

The process of grouping the attributes are described by explaining the Figures 6–10 as follows:

Figure 6 shows the comparison if patients smoke or drink alcohol. In case the patient is a smoker, he is characterized as group 1, and in group 2 if the situation is vice versa. The same situation is followed if the patients is an alcohol drinker than he is characterized in group 1, and in group 2 if he does not.

Figure 7 presents the BP, for which if the patient BP is less than 80, then he is grouped into group 0, and group 1 if it is equal to or more than 80 and group 2 when the blood pressure is higher than 90.



```
If txtBP.Text.Trim().Substring(4, 2) < "80" Then
       strBp = "0"

ElseIf txtBP.Text.Trim().Substring(4, 2)>= "80"
      And txtBP.Text.Trim().Substring(4, 2) <= "90" Then
       strBp = "1"

ElseIf txtBP.Text.Trim().Substring(4, 2) > "90" Then
       strBp = "2"
End If
```

Figure 7. Grouping blood pressure (BP)

Figure 8 demonstrates the grouping of heart rate. If the heart rate is below or equal to 100, it is categorized in group 0 while if it is more than 100 or less then 150, it is categorized as 1. Similarly, in case the heart rate is greater or equal to 150, the patients are categorized in group 2.

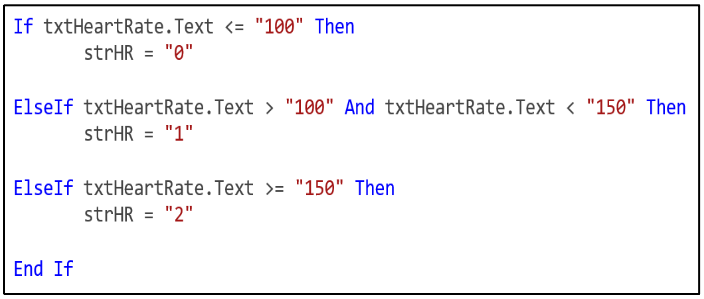
Figure 9 displays the patients grouping as per his cholesterol level. The patient is grouped in 0, when the level of cholesterol is below 5.2, and in group 1 if the value of cholesterol is more or equal to 5.2. Similarly, these are grouped 2, when it is more than or equal to 6.2.

Figure 10 provides a list of conditions for chest pain. As per the program, if the grouping of the chest pain is done as 1, the patient experience chest pain, and 2 in case a risk of cardiac arrest prevails. It is grouped as 3 if a patient is experiencing coronary artery disease. Whereas for being grouped as 4, the patient is at high risk of chest pain.

Lastly, the class comparison results are presented in rules forms (Table 2) and process flow of the system is depicted in Figure 11.

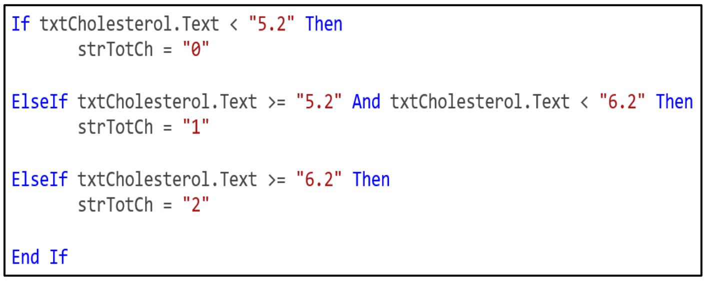
5.4 Testing Results of the Application and its Predictive Performance

As explained above, the resulting conditions provide a comparison of the prime target classes and the executed "if condition rule" (Table 2), where risk results are the outcome of applying the aforementioned rules.



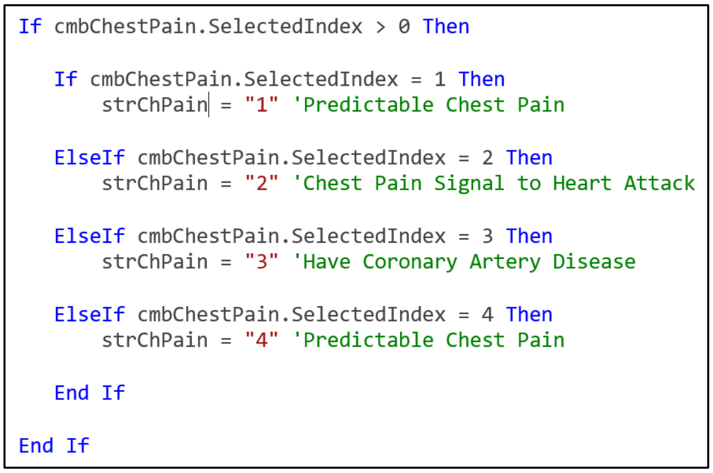
```
If txtHeartRate.Text <= "100" Then
      strHR = "0"

ElseIf txtHeartRate.Text > "100" And txtHeartRate.Text < "150" Then
      strHR = "1"

ElseIf txtHeartRate.Text >= "150" Then
      strHR = "2"

End If
```

Figure 8. Grouping heart rate



```
If txtCholesterol.Text < "5.2" Then
      strTotCh = "0"

ElseIf txtCholesterol.Text >= "5.2" And txtCholesterol.Text < "6.2" Then
      strTotCh = "1"

ElseIf txtCholesterol.Text >= "6.2" Then
      strTotCh = "2"

End If
```

Figure 9. Grouping cholesterol



```
If cmbChestPain.SelectedIndex > 0 Then

   If cmbChestPain.SelectedIndex = 1 Then
       strChPain = "1" 'Predictable Chest Pain

   ElseIf cmbChestPain.SelectedIndex = 2 Then
       strChPain = "2" 'Chest Pain Signal to Heart Attack

   ElseIf cmbChestPain.SelectedIndex = 3 Then
       strChPain = "3" 'Have Coronary Artery Disease

   ElseIf cmbChestPain.SelectedIndex = 4 Then
       strChPain = "4" 'Predictable Chest Pain

   End If

End If
```

Figure 10. Grouping chest pain

The software application is implemented using VB, and the samples of the screenshots are depicted in Figures 12 to 15. These screenshots serve as evidence for the program utility concerning the NCDs prediction for the tested patients. These samples of NCD predictions are shown for different risk levels: no risk, low risk, medium risk, and high risk.

5.5 Findings

The results are examined on the basis of the response gathered through a questionnaire survey, which was held among the doctors and the results achieved from the use of the developed application.

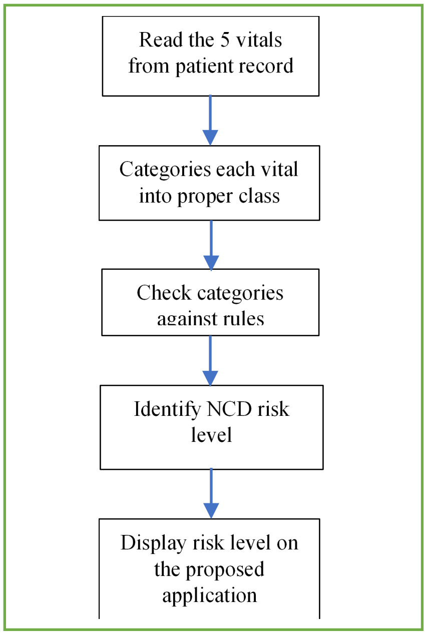
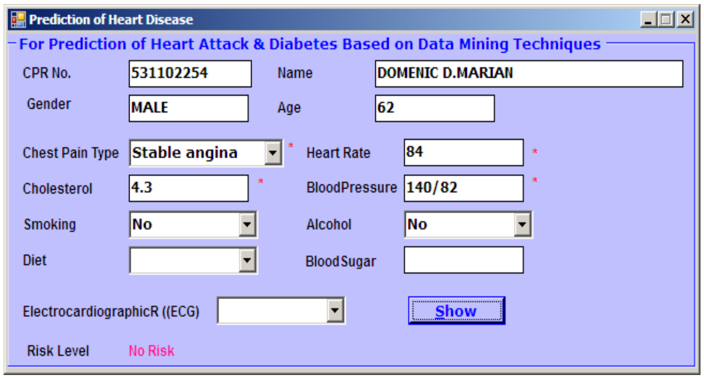

Read the 5 vitals from patient record
Categories each vital into proper class
Check categories against rules
Identify NCD risk level
Display risk level on the proposed application

Figure 11. Flow diagram of the proposed application

Table 2. Classes and Risk Results Comparison

BP	Heart Rate	Cholesterol	Chest Pain	Age	Risk Result
0	0	0	0	0	No risk
0	1	0	1	0	No risk
0	2	0	1	0	No risk
1	1	1	1	1	Low risk
1	2	0	1	0	No risk
1	2	0	1	1	No risk
1	2	0	1	2	Low risk
1	2	0	2	2	Medium risk
1	2	1	1	1	Low risk
1	2	1	2	1	Medium risk
2	2	1	1	1	Low risk
2	2	1	1	2	Medium risk
2	2	1	2 or 3	2	High risk
2	2	2	1	1	Medium risk
2	2	2	2 or 3	2	High risk



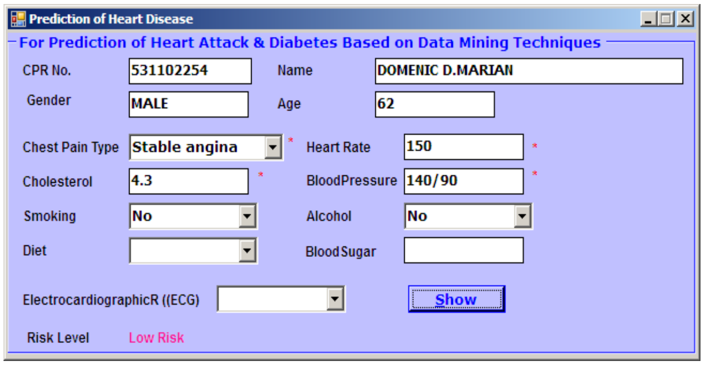
For Prediction of Heart Attack & Diabetes Based on Data Mining Techniques

CPR No. 531102254 Name DOMENIC D.MARIAN
Gender MALE Age 62
Chest Pain Type Stable angina * Heart Rate 84 *
Cholesterol 4.3 * BloodPressure 140/82 *
Smoking No Alcohol No
Diet BloodSugar
ElectrocardiographicR ((ECG)
Risk Level No Risk

Figure 12. Sample of NCD prediction level: no risk

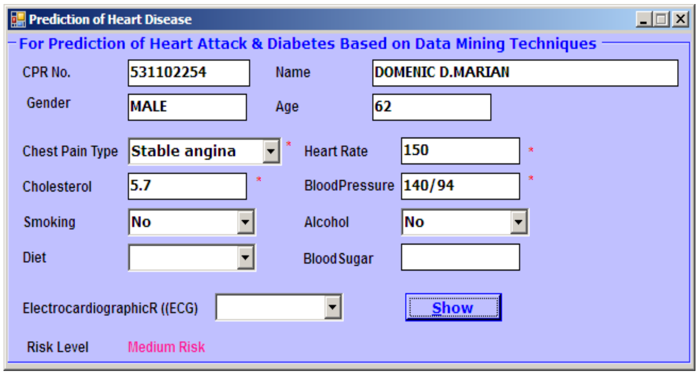
A. Survey Based Questionnaire Inferences

The questionnaire survey provides evidence that at present, there are no tools or the mechanism available in the BDFH, which is having almost 23,000 registered patients, for NCD forecasting and prediction. The patient's record provides a massive volume data, yet there is no examination or the analysis of such essential data for the purpose of forecasting. The survey highlighted that most of the doctors are aware regarding the tools which are being used somewhere else and they will like to have applications that can help them better in the prediction of a patient's possibility for the occurrence or the recurrence of NCDs.


Prediction of Heart Disease
For Prediction of Heart Attack & Diabetes Based on Data Mining Techniques
Show

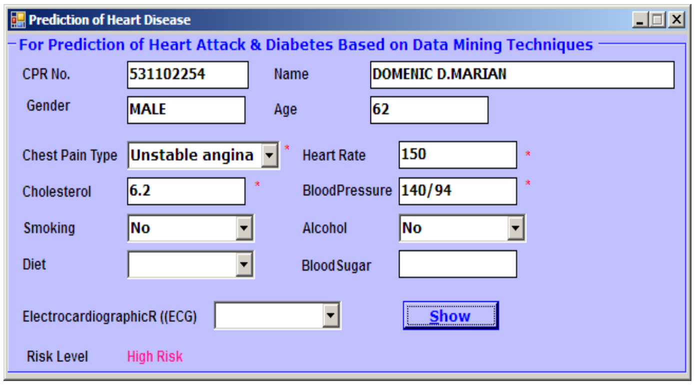
CPR No.	531102254	Name	DOMENIC D.MARIAN
Gender	MALE	Age	62
Chest Pain Type	Stable angina *	Heart Rate	150 *
Cholesterol	4.3 *	BloodPressure	140/90 *
Smoking	No	Alcohol	No
Diet		Blood Sugar	
ElectrocardiographicR ((ECG)			
Risk Level	Low Risk		

Figure 13. NCD prediction level (sample): low risk


Prediction of Heart Disease
For Prediction of Heart Attack & Diabetes Based on Data Mining Techniques
Show

CPR No.	531102254	Name	DOMENIC D.MARIAN
Gender	MALE	Age	62
Chest Pain Type	Stable angina *	Heart Rate	150 *
Cholesterol	5.7 *	BloodPressure	140/94 *
Smoking	No	Alcohol	No
Diet		Blood Sugar	
ElectrocardiographicR ((ECG)			
Risk Level	Medium Risk		

Figure 14. NCD prediction level (sample): medium risk


Prediction of Heart Disease
For Prediction of Heart Attack & Diabetes Based on Data Mining Techniques
Show

CPR No.	531102254	Name	DOMENIC D.MARIAN
Gender	MALE	Age	62
Chest Pain Type	Unstable angina *	Heart Rate	150 *
Cholesterol	6.2 *	BloodPressure	140/94 *
Smoking	No	Alcohol	No
Diet		Blood Sugar	
ElectrocardiographicR ((ECG)			
Risk Level	High Risk		

Figure 15. NCD prediction level (sample): high risk

B. Analysis of the developed software Inferences

The results obtained were discussed with the doctors and are considered as the levels of risks. The doctors were highly contented with the results as they accurately matched their expectations regarding the levels of risk. All kinds of risks are categorized appropriately according to the formerly mentioned categories. Additionally, they indicated high interests in the implementation of such applications live in their clinics. The application of the software has a proof for the accuracy and the efficacy of the suggested technique of the data mining for the forecasting of the NCDs for the patients of BDFH. However, the NCDs can lead to the possible mortality among the patients, the doctors who had a trial of the software feel excitement for its application as it can help them in the detection. As a matter of fact, they want to use it as soon as possible because they highly noticed that it supports quickly and gives better forecasting along with the briefing of the patient's record. On the other hand, the statistical analysis is not applicable to this study as the results are evaluated manually by the consulted practitioners.

Using the application, more savings in medical expenditure can occur avoiding loss of the duty timings and maximizing the utilization of the crucial medical facilities. This application will give the one-stop-shop to all the patients to get the accurate and the calculated information regarding their health. This application is considered to be an asset for the doctors so that they make sure that their detection or inferences are professional as well as correct. Further, such information can be utilized globally, for the patients having the health problems and have to travel abroad, the medical history and the forecasting for the NCD can be made online available or can be shared via emails.

6. Conclusions and Future Work

An application which uses the data mining algorithm for the class comparison has been invented to forecast the level of risk of occurrence or recurrence of NCDs such as diabetes and the heart diseases. Additionally, the outcomes of the application highlighted that the forecasting system could forecast NCDs efficiently, instantly, and effectively. This application helps the doctors to make the decisions regarding the health risks of a patient. It also creates the results, which make it closer to the situations of real life. Data mining, therefore, is more supportive for the health sector and is essential for exploring the knowledge to be used in the health care sector.

This research confirms that the health centers should have a software application or the system which can forecast the NCDs. To have raw data in a large volume is a must for the proper categorization rules which lead to the proper forecasting of NCDs. The forecasting is helpful for not only the physicians but for the patients too for their constant health check-ups. The application developed must be trialed on a constant basis, and the coding must be accurate in order to adopt the application across the various health centers and the clinics. Further, in the future, it can also be considered for the development of the mobile application for the patients where they can input their vitals and get the results instantly.

The developed application and the software can be extended for assisting in forecasting the level of risks for the other NCDs, for example, cancer, Alzheimer, chronic lung disease, stroke, and the osteoporosis. This can be achieved more by highlighting the associated vitals and their analysis to set the appropriate rules by the consultation linked with the expert practitioners. Finally, the obtained results can be further analyzed once the approach is applied on predefined data set with previously known level of risk of each NCD disease.

References

[1] Anon., 2016. World Health Day 2016. [Online] Available at: http://www.who.int/diabetes/global-report/WHD16-press-release-EN_3.pdf?ua=1.

[2] W. H. Organization, "The top 10 causes of death," 24 May 2018. [Online]. Available: http://fmrglobalhealth.com/frame/top10.html. [Accessed 28 June 2019].

[3] Anon., 2018. Diabetes country profiles. [Online] Available at: https://www.who.int/diabetes/country-profiles/bhr_en.pdf?ua=1 [Accessed 28 June 2019].

[4] Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P., 1996. From data mining to knowledge discovery. In Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, pp. 495–515.

[5] Clifton, C., 2010. Encyclopaedia Britannica: definition of data mining.

[6] Obenshain, M. K., 2004. Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology, 25(8), pp. 690–695. DOI: https://doi.org/10.1086/502460

[7] Krishnaiah, V., Narsimha, G. & Chandra, S., 2016. Heart disease prediction system using data mining techniques and intelligent fuzzy approach: a review. International Journal of Computer Applications, 136(2), pp. 43–51. DOI: 10.5120/ijca2016908409

[8] Yoo, I. et al., 2012. Data mining in healthcare and biomedicine: a survey of the Literature. Journal of Medical Systems, 36, pp. 2431–2448. DOI: https://doi.org/10.1007/s10916-011-9710-5

[9] Han, J., Pei, J. & Kamber, M., 2011. Data mining: concepts and techniques. Walthman, MA: Morgan Kaufman Publisher. DOI: 10.1007/978-1-4419-1482-6_3752

[10] Benneyan, J. C., Lloyd, R. & Plsek, P., 2003. Statistical process control as a tool for research and healthcare improvement. Quality & Safety in Health Care, 12, pp. 458–464. DOI: 10.1136/qhc.12.6.458

[11] Mena, J., 2011. Machine learning forensics for law enforcement, security and intelligence. In CRC Press.

[12] Campos, J. et al., 2017. A big data analytical architecture for the asset management, Elsevier, pp. 369–374. https://doi.org/10.1016/j.procir.2017.03.019

[13] Piatetsky-Shapiro, G. & Parker, G., 2011. Lesson: data mining, knowledge discovery: an introduction. In Introduction to Data Mining.

[14] Maimon, O. & Rokach, L., 2011. Data mining and knowledge discovery handbook, Springer.

[15] Hersh, W. R. & Hoyt, R. E., 2018. Health informatics: practical guide seventh edition. lulu.com.

[16] Athina, L. A. & Konstantinos, S. M., 2008. Handbook of research on distributed medical informatics and e-health. Medical Information Science Reference.

[17] Bronzino, D., 2006. Medical devices and systems, CRC Press.

[18] Acharya, R. U. & Yu, W., 2010. Data mining techniques in medical informatics. The Open Medical Informatics Journal, 4, pp. 20–21. DOI: 10.2174/1874431101004020021

[19] Khaing, H. W., 2011. Data mining based fragmentation and prediction of medical data. Shanghai, pp. 480–485. DOI**:** 10.1109/ICCRD.2011.5764179

[20] Khanmohammadi, S., Adibeig, N. & Shanehbandy, S., 2017. An improved overlapping k-means clustering method for medical applications. Expert Systems With Applications, 67, pp. 12–18. DOI: 10.1016/j.eswa.2016.09.025

[21] Jiang, H., Ching, W. K. & Hou, W., 2016. On orthogonal feature extraction model with applications in medical prognosis, Applied Mathematical Modelling. 40, pp. 8766–8776. DOI: 10.1016/j.apm.2016.05.011

[22] Suvarna, C., Sali, A. & Salmani, S., 2017. Efficient heart disease prediction system using optimization technique, Erode, India, pp. 374–279. DOI: 10.1109/ICCMC.2017.8282712

[23] Rairikar, A. et al., 2017. Heart disease prediction using data mining techniques, Coimbatore, India, pp. 1–8. DOI: 10.1109/I2C2.2017.8321771

[24] Tomczak, J. & Zieba, M., 2015. Probabilistic combination of classification rules and its application to medical diagnosis. Machine Learning, 101, p. 105–135. doi: 10.1007/s10994-015-5508-x

[25] Zhu, Y. & Fang, J., 2016. Logistic regression-based trichotomous classification tree and its application in medical diagnosis. Medical Decision Making, 36(8), pp. 973–989. DOI: 10.1177/0272989X15618658

[26] Jaganathan, P. & Kuppuchamy, R., 2013. Threshold fuzzy entropy based feature selection for medical database classification. Computers in Biology and Medicine, 43(12), pp. 2222–2229. DOI: 10.1016/j.compbiomed.2013.10.016