# Applying Machine Learning Models to Electronic Health Records for Chronic Disease Diagnosis in Kuwait

Talal M. Alenezi[1]*, Taiseer H. Sulaiman[2], Amr M. AbdelAziz[3]
Faculty of Computers and Information, Assiut University, Assiut, Egypt[1]
Information Science Dept.-Faculty of Computers and Information, Assiut University, Assiut, Egypt[2]
Faculty of Computers and AI, Beni-Suef University, Beni-Suef, Egypt[3]

***Abstract*—The leading cause of death nowadays is chronic disease. As a result, personal wellbeing has received a considerable boost as a healthcare preventative strategy. A notable development in data-driven healthcare technology is the creation of a prediction model for chronic diseases. In this situation, computational intelligence is used to analyze electronic health data to provide clinicians with knowledge that will help them make more informed decisions about prognoses or therapies. In this study, various classification algorithms have been implemented namely, Decision Tree, K-Nearest Neighbors, Logistic Regression, Multilayer Perceptron, Naïve Bayes, Random Forest, and Support Vector Machines to examine the medical records of patients in Kuwait who had chronic conditions. For predicting diabetes, the support vector machines classifier was the best classifier for predicting diabetes and kidney chronic diseases. For diabetes, it achieved 88.5% accuracy, and 93.6% f1-score, while for kidney; it achieved 94.9% and 92.6% accuracy and f1-score respectively. For predicting heart disease, MLP was the best and achieved 84.7%, and 87.8% accuracy and f1-score respectively.**

***Keywords*—Chronic diseases; Electronic Health Records (EHR); machine learning; classification**

## I. Introduction

Healthcare systems and governments from all around the world are very interested in finding ways to improve healthcare outcomes while lowering costs. Accurate patient risk assessments from both patients and clinicians are essential to achieving this improvement. The usage of electronic health records (EHRs) for digitally recording patient healthcare contacts has increased significantly in tandem with this focus on improving outcomes (Hsiao et al., 2014). As a result, disease risk prediction technologies that may use EHR data to evaluate patient risk have drawn a lot of interest [1].

There is now a lot of interest in applying machine learning techniques to create disease risk prediction tools from EHR data since machine learning algorithms have experienced significant growth in use. “What is the distribution of accuracy with which all EHR-coded medical occurrences may be anticipated?” is a logical place to start. We provide an answer to this query for coded diagnoses since it has not previously been addressed [2].

By using computational intelligence techniques, this study was able to extract pertinent information from electronic medical records that might be used by doctors to deliver effective care. Electronic health records (EHRs) are instantaneous, patient-centered records with secure access for authorized individuals. EHRs are distinguished by the capability of authorized clinicians to generate and manage health data in digital format, which may be shared with other doctors across various healthcare organizations. The Arabic region faces significant challenges due to chronic ailments. Chronic disease-related mortality has been linked to a sizable number of deaths, according to Kuwait's Ministry of Health [3]. Their statistics show that 41% of fatalities were attributable to heart disease, 15% to cancer, 3% to respiratory problems, and 3% to diabetes. Additionally, according to Kuwait's national plan report on chronic diseases for the years 2017 to 2025, cardiovascular disorders are the main cause of mortality there [4].

The goal of this study is to create a predictive analytics model that can predict the onset of three common chronic diseases diabetes, heart disease, and kidney disease early on using electronic health records that contain important information about patients from their hospital visits. A key component of prevention is prediction, which enables healthcare organizations to give top priority to patients with the greatest needs and make the best use of their scarce resources.

To attain the research objectives, the authors conducted a study that involved defining the methodology, study population, study sample, study tools, verifying the validity and reliability of the tools, and utilizing statistical treatment in analyzing the results. The study population comprised all healthcare stakeholders in Kuwait hospitals, and the study sample consisted of (30) medical sector workers, from which a random sample was selected. The study sample's frequencies and percentages were computed, and they are represented in the basic data that includes Gender: (43% males and 56.7% females), Age: (26.7% less than a year, 43.3% between 30 to 40 years, 16.7% between 40 to 50 years, and 13.3% over 50 years), Position: (16.7% manager, 40% attending physician, and 43.3% technician), Number of years of experience (33.3% less than three years, 6.7% between three years to less than six years, and 60% more than six years), and Computer skills: (3.3% weak, 36.7% medium, 60% good). The paper evaluates physicians' perceptions on the current management of chronic diseases in primary healthcare centers using electronic health

record systems, shedding light on the practical implications of using EHR for disease management.

From Table V in Appendix A, it can be observed that the general average for the first dimension, technological infrastructure environment, had a response rate (High), an arithmetic means of (2.66), a standard deviation of (.409), and phrase No. (27) appeared in the first order. (The need to employ e-health solutions to improve the quality of the healthcare sector.) Phrase No. 22 came in second place (There is a need to establish a well-designed primary health care network) with an arithmetic mean (2.87), a standard deviation (.346), and a response rate (High), and it was followed in the last position by phrase No. 23 (There is a need for the establishment of a well-designed primary health care network). (25) (Health informatics is already implemented in the State of Kuwait.) with an arithmetic mean (2.43), and a standard deviation (.679) with a response score (High), and the rest of the expressions came with a response score (High), The standard deviations for the expressions in the first dimension varying between (.681 - .305), which are regarded as low values. This implies that the study sample's reactions to these assertions were consistent. The emergence of the technical infrastructure environment can be explained by the fact that there are numerous opportunities for developing information systems in Kuwaiti hospitals, in addition to the possibility of directing a large investment towards developing the state's infrastructure relying on electronic systems in facilitating work in the medical field, beginning with data recording, and progressing to electronic solutions.

Table VI in Appendix B shows that the overall average for the second dimension, administration, and organization, had a (High) response rate, an arithmetic mean of (2.62), and a standard deviation of (.473). Phrase 31, (The need for high-quality services in the public and private sectors of healthcare.) came in first place with an arithmetic mean (2.93), a standard deviation (.254), and a response score (High), and phrase 32 came in second place (The need for a national health information management strategy to guide public health policies.) Assuming an arithmetic mean (2.27), The rest of the expressions received a response score (High) and a standard deviation (.828), while the standard deviations for the expressions in the second dimension varying between (.254-.828), indicating homogeneity of opinions of the study sample towards these statements. The need to establish a national plan for health information management to lead public health policies might explain the appearance of the administration and organizations with a high level of responsiveness. It also mentioned the need to enhance the quality of services in the field of health care in general, by focusing on health insurance and trying to reduce health expenditures. The occurrence of phrase No. (34) in the last order can be explained by the fact that, despite the availability of an integrated framework for information and technology, this framework is not primarily relied upon in the field of medical care in Kuwait.

The previous study's recommendations emphasized the need of using electronic medical records to register patients' health information rather than paper records. Furthermore, the study emphasized the need for predictive analytics models that can anticipate the emergence of chronic illnesses at an early stage utilizing electronic health information. The study concentrated on diabetes, heart disease, and chronic kidney disease. Heart disease is a major cause of death in the United States; it kills 610,000 of people each year, accounting for one in every four fatalities. [5]. Diabetes is recognized as a chronic illness with the world's fastest growing rate at the same time [6]. Diabetes affects 415 million people worldwide, accounting for 12% of total medical spending ($673 billion) [6]. Diabetes affected 29.1 million Americans in 2012, accounting for around 9.3% of the population [6]. About 700 million people worldwide are affected with Chronic Kidney Disease (CKD) each year, resulting in the deaths of almost 1.2 million people [7]. Cardiovascular diseases such as myocardial infarction, stroke, and heart failure are more prevalent in chronic renal disease patients [7]. Patients who have both illnesses have a worse prognosis [8].

The following are the primary contributions of the article:

- Using electronic health records instead of paper records.
- Using accessible datasets from patients' medical records, machine learning techniques are used to predict the existence of chronic illnesses.
- Examining medical records of all patients to ensure proper diagnosis of chronic disorders.
- Identifying new patients with comparable symptoms and illness development phases based on physician supervision and medical record analysis for a specific type of chronic disease.

The second half of the paper will go into similar research, while Section III will examine the datasets used in the study. The fourth part will offer a full description of the suggested technique. Following that, Section V will show the test findings and assess the proposed strategy. Finally, Section VI will give findings and recommendations for further research.

## II. Related Work

Yaser et al. [9], has presented a research paper. The fundamental contribution of this work is the development of a medical recommendation system that employs the closest neighbor’s classification approach and the collaborative filtering methodology. The suggested strategy was evaluated based on two criteria: statistical correctness and efficacy in giving patient counseling. The results showed that the suggested method outperformed earlier approaches in diagnosing patients, which was ascribed to the use of the "close neighbors" strategy. The suggested technique exhibited great accuracy in patient diagnosis and gave suitable therapy recommendations. However, due to the restrictions of patient data privacy, getting exact hospital data remains a substantial barrier for the suggested strategy.

Chicco et al. [10], presented another paper. A dataset of 491 patients from the United Arab Emirates was analysed to find independent risk variables for CKD at stages 3-5. The authors utilised two strategies, one based on classic univariate biostatistics testing and the other on machine learning. The biostatistics tests revealed that 68.42% of clinical parameters were significant but lacked accuracy. As a result, the authors

used Random Forests to determine feature ranking. The study proved that computational intelligence could predict significant CKD development with or without time information, and that the most important clinical variables alter when the temporal component is incorporated. The benefit of this study is that it provides a full examination of risk variables connected with CKD at stages 3-5 and applies machine learning techniques to determine the most essential clinical characteristics. However, the scientists did not examine the therapeutic significance of the findings, instead focused on developing and strengthening computational intelligence technologies.

Hohman et al. [11], presented The MENDS project which has successfully proved its capacity to satisfy principles and criteria aimed at optimising the project and supporting the CDC's DMI. The project team used statistical approaches to obtain credible prevalence estimates from EHR data that appropriately represent the underlying populations at geographic target levels. The team was able to generate more reliable local, state, and national prevalence estimates by using the breadth of clinical data. The initiative also provided an opportunity for the DOH to expand its EHR-based surveillance beyond federal programme requirements. However, obstacles such as outreach, communication, paperwork, research, and training have hindered the project's timeframe. The team is also comparing prevalence figures from other national chronic illness surveillance systems to determine their validity. The capacity to combine standard public health monitoring techniques with EHR-based surveillance and gain relevant insights from the data is one of the project's advantages. The drawbacks include the technical work and expense necessary to comply with federal rules, as well as the additional outreach, communication, documentation, investigation, or training that is required.

Kumari and Seema [12], used four distinct classifiers to detect diabetes and heart disease: Nave Bayes, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). The results showed that SVM had the best accuracy rate for heart disease at 95.556%, while Nave Bayes had the highest accuracy rate for diabetes at 73.588%. The benefit of employing these classifiers is that they can help clinicians find successful therapies and best practices, resulting in better and more direct healthcare for patients. However, these classifiers have the disadvantage of requiring a significant quantity of data to reach accuracy and can be time-consuming to train.

Perotte et al. [13], have evaluated the usefulness of employing electronic healthcare data and ICD-10 pain-related diagnostic clusters to self-reported data, which is considered the gold standard, in identifying persons with Musculoskeletal Chronic Pain (MSCP). The study had a 61% response rate, and the findings indicated that the prevalence of MSCP was somewhat lower when detected by electronic health records (14.7% weighted prevalence) than when found with survey questionnaires (25.9% weighted prevalence). The study found a poor level of agreement between the two MSCP categories (kappa agreement statistic = 0.21). When compared to MSCP determined using self-report as the benchmark, using electronic health data exhibited a sensitivity of 30.9%, specificity of 91.0%, and positive predictive value of 54.5%. When age and gender were taken into consideration, patients with MSCP identified through electronic health records or self-report had higher levels of pain-related disability, pain severity, depressive symptoms, and long-term opioid use, compared to individuals with single-site chronic pain identified through the same method. This study demonstrates the use of electronic health records in identifying people with MSCP, potentially leading to a greater detection of people with significant chronic pain in studies that are population-based. The study's weaknesses, however, include a low survey response rate and a lack of agreement between the two MSCP categories.

Hazazi and Wilson [14] aimed to evaluate physicians' perceptions on the current management of Noncommunicable Diseases (NCDs) in Primary Healthcare Centers (PHCs), emphasising on the role of the Electronic Health Record (EHR) system, according to a separate article provided by. The study included semi-structured interviews with 22 Ministry of Health physicians who worked in chronic illness clinics at PHCs. While physicians recognised the benefits of using EHRs, such as improved accuracy in patient documentation and access to patient information, they also identified areas for improvement, such as the lack of a patient portal for patients to access their health information and the system's inability to facilitate multidisciplinary care. In general, doctors viewed the EHR system favorably, although its influence on patient care at chronic illness clinics remained limited.

Areej and Malibari [15] used EO-LWAMCNet, MSSO-ANFIS, T-RNN, and DLMNN algorithms to predict heart and kidney illnesses in real time. On two separate datasets, the suggested model has an accuracy of 93.5% and 94%. The suggested model's merits include its capacity to forecast chronic illnesses including heart and renal disease in real time, its optimal performance, and its short execution time. The suggested model, however, has numerous shortcomings, notably its high cost and inability to distinguish between positive and negative predictions.

## III. DATA

In this study, an electronic healthcare record (EHR) dataset comprising information about patients in Kuwaiti hospitals was constructed. Each row in the dataset represents a single patient, and the columns indicate all the patient's attributes/features, as detailed in Table I. This dataset was created for the purpose of predicting diabetes, heart disease, and chronic kidney disease by combining all characteristics from original datasets into a single dataset file for use in training and testing prediction models, in addition to the personal information of all patients during all hospital visits. This EHR dataset will be published soon after the government approval. Diabetes illness data was obtained from the Irvine's Center for Machine Learning and Intelligent Systems, California University [16]. Rather than more than 60,000 individual patients, this study includes clinical data from over 100,000 unique interactions. The information was gathered from roughly 74 million patient visits involving 17 million people [16]. This study's data was gathered over a ten-year period, from 1999 to 2008, and includes many credits that correspond to the seasons of confirmation and release for diabetes patients. The records

include information on laboratory tests and procedures, diagnoses, and drugs given during the hospitalization.

The dataset utilized in this work for heart disease prediction was collected from the UCI Machine Learning Repository of Irvine C.A, University of California and was freely accessible online [17]. The cardiac disease datasets utilized in this study have identical properties and example designs. Only 14 of the 76 raw qualities in these datasets are considered essential, including anticipated property. The Cleveland Clinic Foundation dataset has 303 patient records, whereas the Hungarian Institute of Cardiology dataset contains 294 patient records. This study examined a dataset of 491 individuals from the United Arab Emirates published by Al-Shamsi et al. for chronic renal disease [18]. The data for this study was gathered at Tawam Hospital in Al-Ain (Abu Dhabi, United Arab Emirates) in 2008. The dataset offered 491 patients, 241 women and 250 males, with an average age of 53.2 years. Each patient was given a chart with 13 clinical variables that represented the results of laboratory tests and examinations as well as data from the dataset. The authors used multivariable Cox proportional hazards to identify the risk variables that cause CKD at stages 3-5. However, the analysis did not include a prediction phase, which may have recovered more relevant information or previously unknown patterns in the data.

TABLE I. NAME AND TYPE OF EACH FEATURE OF THE EHR DATASET

| Feature Name | Type | Feature Name | Type |
|---|---|---|---|
| Encounter ID | Numeric | Examide | Numeric |
| Patient number | Nominal | Insulin | Numeric |
| Race | Nominal | Anemia | Nominal |
| Gender | Numeric | Creatinine hosphokinase | Nominal |
| Age | Numeric | Blood pressure | Numeric |
| Weight | Nominal | Serum creatinine | Numeric |
| Admission type | Nominal | Smoking | Nominal |
| Discharge disposition | Numeric | Specific gravity | Nominal |
| Admission source | Numeric | Albumin | Numeric |
| Time in hospital | Nominal | Red blood cells | Numeric |
| Payer code | Nominal | Pus cells | Nominal |
| Medical specialty | Numeric | Bacteria | Nominal |
| Number of lab procedures | Numeric | Blood Urea | Numeric |
| Number of procedures | Numeric | Sodium | Numeric |
| Number of medications | Numeric | Potassium | Numeric |
| Number of outpatient visits | Numeric | Hemoglobin | Numeric |
| Number of emergency visits | Numeric | White Blood Cell Count | Numeric |
| Number of inpatient visits | Nominal | Hypertension | Numeric |
| Diagnoses 1 | Nominal | CoronaryArtery Disease | Nominal |
| Diagnoses 2 | Nominal | Appetite | Nominal |
| Diagnoses 3 | Numeric | A1c test result | Nominal |
| Number of diagnoses | Nominal | Change of medications | Nominal |
| Glucose serum test result | Nominal | Readmitted | Nominal |

## IV. METHODOLOGY

The task stated previously is to apply binary classification to predict the existence of diabetes, heart, and kidney chronic diseases using the data said earlier. This part will display the classification models applied in this study. The proposed method architecture is described in Fig. 1.

### A. Data Acquisition

In this step, we gather electronic health record (EHR) information from medical facilities, healthcare centers, or publicly available datasets, while guaranteeing that the data is anonymized and compiles with applicable privacy regulations.

### B. Data Preprocessing

Before we use the datasets for prediction, we analyze them to obtain their drawbacks. In the diabetes datasets, we observed that there are many features with missing value ratio, as shown in Fig. 2.

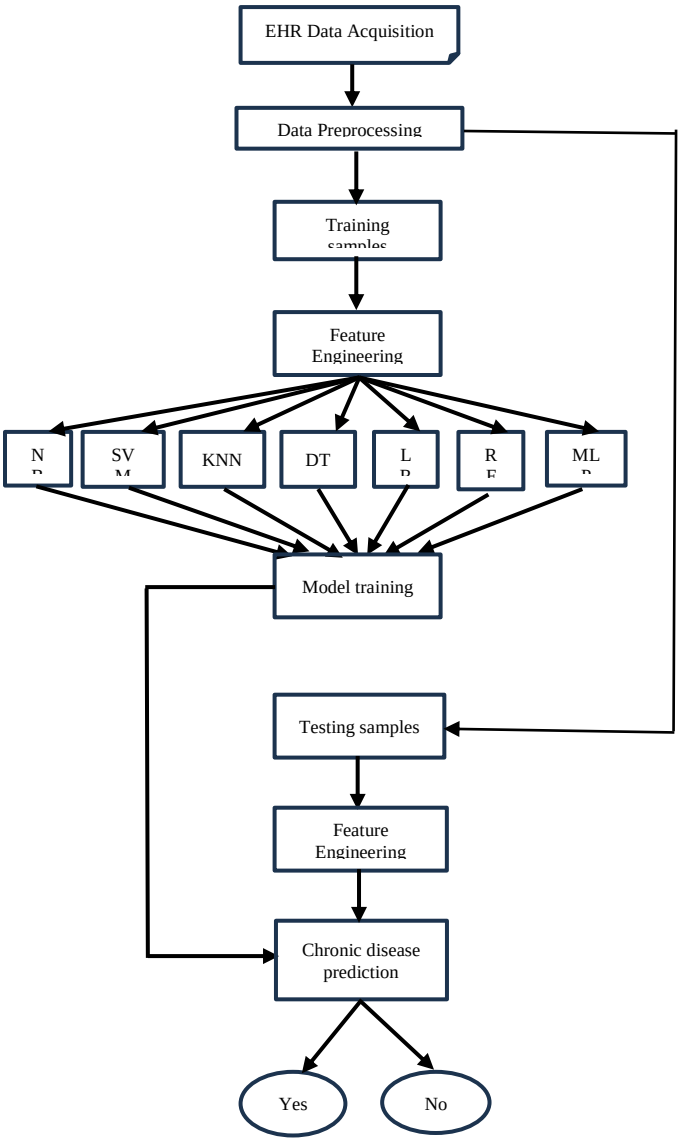



Fig. 1. Proposed method architecture.

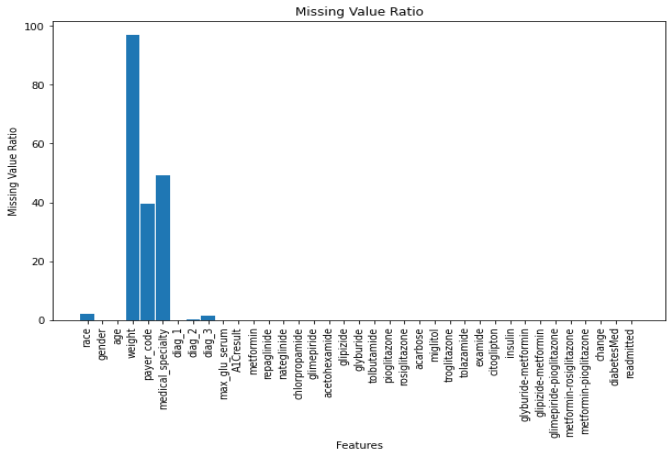



Fig. 2. Features with missing value ratio in diabetes dataset.

After applying feature reduction based on missing value ratio with ratio more than or equal 40%, "weight", "payer_code" and "medical_specialty" features have been dropped from dataset. Then we are dealing with features that have less than 40%, we have "diag_1", "diag_2", "diag_3" and "race" features. We have removed any row that have a missing value in any features and this through the following:

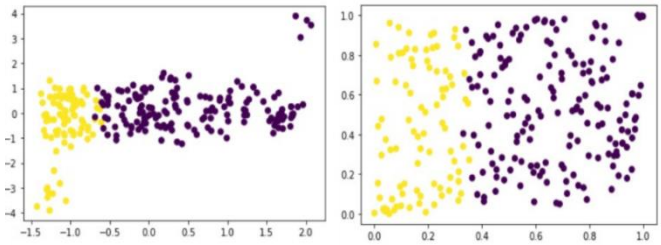
- Retrieving the index of the row where all three (diag_1, diag_2 and diag_3) values are missing and storing it in a set.
- Retrieving the index of the rows where only diag_1 value is missing and appending the same in the set.
- Retrieving the index of the rows where only diag_2 value is missing and appending the same in the set.
- Retrieving the index of the rows where only diag_3 value is missing and appending the same in the set.
- Retrieving the index of the rows where only race value is missing and appending the same in the set.
- Then drop all rows that have an index in the set.

Now we have a dataset without missing value. The next step is to remove columns (features) that have no relation with our output; in another word "featureless" such as "encounter_id" and patient_nbr".

So, we now have a clean dataset without any missing value and featureless columns but there are some columns still with text not numbers. In order to that, we have labeled these columns with integer numbers. Then to make our classifiers and model avoid overfitting, we make data regularization to all data.

Pearson's Correlation Coefficient helped us identify the connection between two quantities by assessing the intensity of their association. Also, it provides a metric ranging from -1 to +1 to quantify this relationship. A value of one signifies a strong correlation, while zero indicates no correlation. A heatmap is a visual representation of data in two dimensions, utilizing color to convey information. This graphical method assists users in visualizing both simple and complex data. This heat map is shown in Fig. 4 in Appendix C below. We observed the last row, "target" and noted its correlation scores with different features. "examide", "citoglipton" and "metformin-rosiglitazone" features are not correlated with "target". They don't contribute much to the model, so we dropped them.

To treat the outliers, we used Quantile Transformer. This technique converts the features into a uniform or normal distribution. Therefore, this conversion typically causes the most common values of a feature to be more widely distributed. Furthermore, this preprocessing scheme is considered robust as it effectively lessens the influence of minor outliers in the dataset. Fig. 3 shows the difference between applying Standard Scalar vs. Quantile Transformation. In Standard Scalar, Fig. 3(a), the Y-axis has eight units, whereas X-axis has only 3.5 units, indicating that Outliers have affected the scales. After applying Quantile Transformation, the Y-axis, and X-axis are equally scaled as shown in Fig. 3(b). The outliers are still present in the dataset, but their impact has been reduced.



(a) (b)

Fig. 3. (a) Standard scalar.(b) Quantile transformation.

We applied these preprocessing processes to heart and kidney chronic diseases datasets as they had the same problems, imbalance, and outliers.

### C. Feature Engineering

Feature engineering involves identifying and extracting meaningful features from electronic health record (EHR) data that may be associated with chronic diseases. This process includes both feature extraction, which involves identifying relevant features, and feature selection, which involves selecting the most informative features for predicting the diseases.

### D. Split Data

After Partitioning the dataset into appropriate training, validation, and testing subsets, using a suitable ratio, as follows, the training set contains 70% of the data, the validation and the test set are 15% each.

### E. Classifiers

*1) Naïve Bayes (NB):* The "Naïve Bayes Classifier" refers to straightforward probability-based classifier which utilizes Bayes' theorem with dependable assumptions of independence. It anticipates that the proximity or lack of a particular class component will depend on the proximity or lack of any other element [19]. Conditional probabilities are necessary for the Naive Bayes calculation. It employs the Bayes hypothesis, a method for computing probability

through frequency counting of different qualities and attributes combinations in the collected data. The Bayes Theorem estimates the likelihood of an event, given the possibility of an earlier event. The Bayes hypothesis can be stated as follows [24]. When "b" addresses the current event for the needy and "a" addresses a previous occasion, Bayes' theorem can be formulated in the following manner:

$$Prob\ (a|b)\ = \frac{Prob\ (b|a) * Prob(b)}{Prob(a)} \qquad (1)$$

Here, Prob $(a|b)$ denotes the probability of "b", and Prob $(b|a)$ represents the probability of "b" given "a". The main benefit of the Naive Bayes classifier is the minimal training data it requires to compute mean and variance parameters essential for classification. Under the assumption of independent variables, Variances of for each class's variables need to be calculated only, rather than the overall variance. The reason for using the naïve bayes is that it is fast, scalable and useful for dealing with missing data.

*2) Support Vector Machine (SVM):* SVM is a supervised learning technique used for regression and classification tasks, known as Support Vector Regression (SVR) and Support Vector Classification (SVC), correspondingly. It is recommended for use with more diminutive datasets as the processing time for larger datasets is lengthy [20]. The fundamental principle of SVM involves determining the hyperplane which efficiently splits the dataset's features to distinct domains, where the points closest to the hyperplane are designated as support vectors, and their distance from the hyperplane is known as the margin. This algorithm aims to locate a hyperplane that minimizes the likelihood of misclassifying cases in the test dataset [21]. SVM is very useful when the data is not distributed regularly, and it doesn't suffer from overfitting.

*3) K-Nearest Neighbor (KNN):* KNN is a supervised learning algorithm that employs proximity to classify or forecast the grouping of a singular data point [22]. Although support vector machines (SVM) can be used for classification and regression tasks, it is commonly used as a classification technique, relying on the concept that analogous points are usually in proximity to one another. KNN is a lazy learning technique that doesn't generalize the training data points, resulting in a quick training phase. It predicts the classification of a new sample point using data from multiple classes. Additionally, KNN is non-parametric, making no assumptions about the data, and the model is derived directly from the data [24]. KNN was used since it doesn't need to be trained before producing predictions, and so new data can be supplied without disrupting the system's accuracy.

*4) Decision tree:* The classification process can be directed using decision trees [23]. They are arranged in a simple tree structure, with terminal nodes displaying the results of decisions and non-terminal nodes representing tests on one or more attributes. By using information gain and gain ratio measures as splitting models, C4.5 improved the ID3 decision tree induction algorithm. The decision tree construction algorithm can be summed up as follows: 1) Decide which attribute best distinguishes the values of the output attribute. 2) Make a distinct branch of the tree for each attribute value. 3) Create subgroups from the instances to correspond with the attribute values of the chosen node. For each subgroup, stop the attribute selection process if the following conditions are met: a) When all of the instances in a subgroup have the same value for an output attribute, the attribute selection process for the current path is stopped, and the branch on the current path is marked with the chosen value. b) There is only one node in the subgroup or no other distinguishing characteristics can be found. Label the branch with the output value seen in most of the remaining cases, just like in (a). 5) Repeat the procedure above for each subgroup created in (3) that has not been designated as terminal [24].

*5) Random forest:* The random forest algorithm, as explained in [25], is composed of multiple decision trees that are trained on a bootstrapped sample of the data set, where one-third of the data is reserved for testing as the out-of-bag sample. To introduce more diversity and reduce correlation among decision trees, feature bagging injects another instance of randomness. Before training, the random forest algorithm requires three crucial hyperparameters to be specified: node size, number of trees, and number of sampled features. The random forest classifier can be employed to address both regression and classification tasks, and the method of determining the prediction varies depending on the type of problem. For a regression task, the decision trees' predictions are averaged, while for classification tasks, the predicted class is determined through a majority vote, selecting the most frequent categorical variable [25].

*6) Logistic Regression (LR):* Logistic regression is a machine learning technique which is categorized as a supervised machine learning model. It is also regarded as a discriminative model that aims to differentiate between classes or categories [25]. Logistic regression is commonly used for prediction and classification problems such as Fraud detection, Disease prediction, and Churn prediction. This analytical approach in medicine can anticipate the probability of disease or illness for a particular population, enabling healthcare organizations to establish preventive care measures for individuals with a higher susceptibility to specific conditions [26].

*7) Multi-Layer Perceptron (MLP):* Three layers make up an MLP, a feed-forward neural network: an input layer, an output layer, and a hidden layer [25]. While the output layer completes various tasks like classification and prediction, the input layer receives the input signal for processing. There are numerous hidden layers that make up the computational engine of the MLP that lie between the input and output layers. Data flows forward from the input layer to the output layer, much like a feed-forward network in an MLP. The MLP neurons are trained with the backpropagation learning algorithm. Since MLPs can approximatively solve any continuous function, they can solve problems that cannot be

linearly separated. They are used in a variety of tasks including classification, pattern recognition, forecasting, and approximation. [27]. We used MLP as it is very flexible and can be used for learning mapping from inputs to outputs generally.

## V. Experimental Results

To evaluate the proposed model, several experiments have been carried out using different parameters as shown in Table II. All experiments have been done using Python programming language by Jupiter notebook editor with some machine learning toolboxes (e.g. Scikit-learn, NumPy, and matplotlib).

TABLE II. Parameters of Some Proposed Models

| Algorithm | Parameters |
|---|---|
| SVM | **Kernel functions** = {linear, sigmoid, RBF} |
| KNN | Number of neighbors ($n$) = {30, 40, 50} |
| MLP | Number of hidden layers ($h$) = {5, 10, 20} |

The accuracy and f1-score are calculated to examine how classification is carried out. As a result of the classifier, four cases are taken into consideration.

- True Positives (TP): The sample size correctly assigned to that class.
- True Negatives (TN): The sample size correctly excluded from that class.
- False Positives (FP): The sample size wrongly excluded from that class.
- False Negatives (FP): The sample size wrongly assigned to that class.

The number of accurate and wrong classifications in each potential prediction of the classified variables is used to evaluate how effective the classification model is. The accuracy is calculated using the following formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

We have utilized three medical datasets related to diabetes, heart, and kidney disease. These datasets were all retrieved from the UCI machine learning library [15, 17, 18]. The objective is to compare all classifiers, including Naïve Bayes, KNN, SVM, DT, RF, LR, and MLP, and to categorize the diseases depending on parameter choices. The accuracy of several classifiers for diabetes and heart disease datasets is shown in Table III. After applying some experiments with different parameters shown in Table II, we decided on the best hyperparameters for SVM, KNN, and MLP classifiers. For SVM, three kernel functions have been utilized and the RBF kernel has achieved the highest accuracy on diabetes and kidney datasets. Three values for the number of neighbors were used for the KNN classifier and the best one was 40. In MLP, the best number of hidden layers was 10, which achieved the highest accuracy on the heart dataset.

In Table IV, a recognized confusion matrix is obtained to estimate four measures, accuracy, precision, recall, and f-score.

TABLE III. Accuracy of All Classifiers for Diabetes, Heart, and Kidney Datasets

| Dataset | Classifier | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|---|
| Diabetes | Naïve Bayes | 88.3% | 85.5% | 78.1% | 83.5% |
| | SVM | **88.5%** | **92.4%** | **99.8%** | **93.6%** |
| | KNN | 88.2% | 57.3% | 51.5% | 49.5 |
| | DT | 82.2% | 91.2% | 85.8% | 84.2% |
| | RF | 88.4% | 44.1% | 51.5% | 47.2% |
| | LR | 88.3% | 69.2% | 52.3% | 48.5% |
| | MLP | 75.1% | 39.5% | 20.4% | 26.4% |
| Heart | Naïve Bayes | 83.1% | 83.6% | 82.5% | 81.6% |
| | SVM | 83.6% | 83.8% | 81.4% | 81.5% |
| | KNN | 83.1% | 83.5% | 80.6% | 81.5% |
| | DT | 80.9% | 79.8% | 78.8% | 78.8% |
| | RF | 84.2% | 84.5% | 84.4% | 85.1% |
| | LR | 82.6% | 82.8% | 83.6% | 84.9% |
| | MLP | **84.7%** | **88.5%** | **87.4%** | **87.8%** |
| Kidney | Naïve Bayes | 86.8% | 82.6% | 77.6% | 74.7% |
| | SVM | **94.9%** | **94.5%** | **90.2%** | **92.6%** |
| | KNN | 81.8% | 85.7% | 53.6% | 54.8% |
| | DT | 89.9% | 88.5% | 68.4% | 82.3% |
| | RF | 86.8% | 84.6% | 71.0% | 75.3% |
| | LR | 85.8% | 87.4% | 67.2% | 71.1% |
| | MLP | 89.8% | 65.7% | 68.5% | 66.4% |

TABLE IV. Confusion Matrix of SVM, DT, and MLP classifiers for Diabetes, Heart, and Kidney chronic datasets

| Dataset | Classifier | Precision | Recall | F-measure | Class |
|---|---|---|---|---|---|
| Diabetes | NB | 0.85 | 0.85 | 0.83 | Yes |
| | | 0.86 | 0.89 | 0.84 | No |
| | SVM | 0.93 | 1.00 | 0.92 | Yes |
| | | 0.92 | 1.00 | 0.93 | No |
| | KNN | 0.25 | 0.04 | 0.06 | Yes |
| | | 0.89 | 0.99 | 0.93 | No |
| | DT | 0.85 | 0.87 | 0.82 | Yes |
| | | 0.97 | 0.83 | 0.86 | No |
| | RF | 0.88 | 1.00 | 0.00 | Yes |
| | | 0.00 | 0.00 | 0.94 | No |
| | LR | 0.50 | 0.01 | 0.03 | Yes |
| | | 0.88 | 1.00 | 0.94 | No |
| | MLP | 0.40 | 0.20 | 0.27 | Yes |
| | | 0.39 | 0.20 | 0.26 | No |
| Heart | NB | 0.85 | 0.87 | 0.86 | Yes |
| | | 0.81 | 0.78 | 0.79 | No |
| | SVM | 0.83 | 0.90 | 0.86 | Yes |
| | | 0.84 | 0.75 | 0.79 | No |
| | KNN | 0.83 | 0.90 | 0.86 | Yes |
| | | 0.84 | 0.74 | 0.79 | No |
| | DT | 0.83 | 0.84 | 0.84 | Yes |
| | | 0.78 | 0.77 | 0.77 | No |
| | RF | 0.85 | 0.89 | 0.87 | Yes |
| | | 0.83 | 0.78 | 0.81 | No |
| | LR | 0.84 | 0.87 | 0.85 | Yes |
| | | 0.81 | 0.77 | 0.79 | No |

| | | | | | |
|---|---|---|---|---|---|
| | MLP | 0.89<br>0.87 | 0.88<br>0.86 | 0.88<br>0.87 | Yes<br>No |
| Kidney | NB | 0.77<br>0.88 | 0.50<br>0.96 | 0.61<br>0.92 | Yes<br>No |
| | SVM | 0.94<br>0.95 | 0.80<br>0.99 | 0.86<br>0.97 | Yes<br>No |
| | KNN | 1.00<br>0.81 | 0.10<br>1.00 | 0.18<br>0.90 | Yes<br>No |
| | DT | 0.86<br>0.91 | 0.60<br>0.97 | 0.71<br>0.94 | Yes<br>No |
| | RF | 0.82<br>0.88 | 0.45<br>0.97 | 0.58<br>0.92 | Yes<br>No |
| | LR | 0.88<br>0.86 | 0.35<br>0.99 | 0.50<br>0.92 | Yes<br>No |
| | MLP | 0.66<br>0.65 | 0.81<br>0.57 | 0.72<br>0.61 | Yes<br>No |

The classification outcomes are presented as a matrix by the confusion matrix. It includes information for actual and predicted classes made using the classification framework. The cell represents the sample size that was classified as true when they were actually true (i.e., TP) and those classified as false while quiet (i.e., TN), respectively. The remaining two cells represent the number of incorrectly classified pieces. In fact, the cells denoting the sample size labeled false when they were actually true (i.e., FN) and the sample size labeled true when they were wrong (i.e., FP) [28]. After the confusion matrices are constructed, the following formulas can determine the precision, recall, and F-measure:

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{3}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{4}$$

$$\text{F-measure} = \frac{2 * TP}{(2 * TP + FP + FN)} \tag{5}$$

## VI. Conclusion and Future Work

In this study, we generated electronic healthcare records with symptoms data sourced from three distinct datasets in the UCI machine learning repository for predicting diabetes, heart disease, and chronic kidney disease. Data of the patients were trained with various classifiers, including Naïve Bayes, SVM, Decision Tree, Random Forest, Logistic Regression, and Multilayer Perceptron. The results of this study indicate that data mining techniques can be effectively employed to identify and predict chronic diseases. The results indicate that the Support Vector Machine (SVM) algorithm was the most successful in predicting diabetes and kidney diseases, while the Multilayer Perceptron (MLP) algorithm was the optimal choice for predicting heart disease. The Naïve Bayes and Random Forest algorithms also demonstrated strong performance across all datasets.

## References

[1] R. Kleiman, P. Bennett, P. Peissig, R. Berg, Z. Kuang, S. Hebbring, M. Caldwell, and D. Page. “High-Throughput Machine Learning from Electronic Health Records”. Proceedings of Machine Learning Research 85:1–24, Machine Learning for Healthcare, 2019.

[2] M. Ghaderzadeh, M. Aria. “Management of Covid-19 Detection Using Artificial Intelligence in 2020 Pandemic”. Proceedings of the 5th International Conference on Medical and Health Informatics. ICMHI 2021, May 14–16, 2021, Japan.

[3] ME. Hossain, A. Khan, MA. Moni, S. Uddin. “Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review”. IEEE/ACM Trans Comput Biol Bioinform. 2021 Mar-Apr;18(2):745-758. doi: 10.1109/TCBB.2019.2937862. Epub 2021 Apr 6. PMID: 31478869.

[4] J. Al-Otaibi, E. Tolma, W. Alali, D. Alhuwail, S. Aljunid. “The Factors Contributing to Physicians’ Current Use of and Satisfaction With Electronic Health Records in Kuwait’s Public Health Care: Cross-sectional Questionnaire Study”. JMIR Med Inform.10(10): e36313 URL: https://medinform.jmir.org/2022/10/e36313 DOI: 10.2196/36313, 2022.

[5] W. Tsao, W. Aday, I. Almarzooq, A. Alonso, Z. Beaton, S. Bittencourt, “Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association”, American Heart Association, https://doi.org/10.1161/CIR.0000000000001052Circulation. 145:e153–e639. 2022.

[6] S. Suh, SO. Song, JH. Kim, H. Cho, WJ. Lee, BW. Lee. “Effectiveness of Vildagliptin in Clinical Practice: Pooled Analysis of Three Korean Observational Studies (the VICTORY Study)”. J Diabetes Res. 2017:5282343. doi: 10.1155/2017/5282343. Epub 2017 Aug 24. PMID: 29057274; PMCID: PMC5613692, 2017.

[7] JC. Lv, LX. Zhang. “Prevalence and Disease Burden of Chronic Kidney Disease”. Adv Exp Med Biol. 1165:3-15. doi: 10.1007/978-981-13-8871-2_1. PMID: 31399958. 2019.

[8] Ammirati, Adriano. “Chronic Kidney Disease. Revista da Associação Médica Brasileira”. 66. s03-s09. 10.1590/1806-9282.66.s1.3. 2020

[9] N, Yaser, L, Zhu, C. Junde, Z, Qiu, X. Yuan, D.N, Yahya, E. Sajad. “Diagnosis of Chronic Diseases Based on Patients’ Health Records in IoT Healthcare Using the Recommender System”. Wireless Communications and Mobile Computing. 2022. 1-14. 10.1155/2022/5663001.

[10] D. Chicco, C. A. Lovejoy and L. Oneto, "A Machine Learning Analysis of Health Records of Patients With Chronic Kidney Disease at Risk of Cardiovascular Disease," in *IEEE Access*, vol. 9, pp. 165132-165144, doi: 10.1109/ACCESS.2021.3133700. 2021.

[11] Hohman, Katherine H. DrPH, MPH; Martinez, Amanda K. MPH, MSN, RN; Klompas, Michael MD, MPH; Kraus, Emily M. PhD, MPH; Li, Wenjun PhD; Carton, Thomas W. PhD, MS; Cocoros, Noelle M. DSc, MPH; Jackson, Sandra L. PhD, MPH; Karras, Bryant Thomas MD; Wiltz, Jennifer L. MD, MPH; Wall, Hilary K. MPH. Leveraging Electronic Health Record Data for Timely Chronic Disease Surveillance: The Multi-State EHR-Based Network for Disease Surveillance. Journal of Public Health Management and Practice 29(2):p 162-173, March/April 2023. | DOI: 10.1097/PHH.0000000000001693.

[12] D. Kumari, S. Seema. “Predictive analytics to prevent and control chronic diseases”. 381-386. 10.1109/ICATCCT.2016.7912028. 2016.

[13] A. Perotte, R. Ranganath, JS. Hirsch, D. Blei, N. Elhadad. “Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis”. J Am Med Inform Assoc. 2015 Jul;22(4):872-80. doi: 10.1093/jamia/ocv024. Epub 2015 Apr 20. PMID: 25896647; PMCID: PMC4482276.

[14] A. Hazazi, A. Wilson. “Leveraging electronic health records to improve management of noncommunicable diseases at primary healthcare centres in Saudi Arabia: a qualitative study”. BMC Fam Pract. 2021 May 27;22(1):106. doi: 10.1186/s12875-021-01456-2. PMID: 34044767; PMCID: PMC8157615.

[15] Areej A. Malibari, An efficient IoT-Artificial intelligence-based disease prediction using lightweight CNN in healthcare system, Measurement: Sensors, Volume 26, 2023, 100695, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2023.100695.

[16] B. Strack, JP. DeShazo, C. Gennings, JL. Olmo, S. Ventura, KL. Cios, JN. Clore. “Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records”. Biomed Res Int.;2014: 781670. doi: 10.1155/2014/781670. Epub 2014 Apr 3. PMID: 24804245; PMCID: PMC3996476.

[17] A. Tanvir, M. Assia, B. Sajjad, A. Muhammad, A. Raza. “Survival analysis of heart failure patients: A case study”. PLOS ONE. 12. e0181001. 10.1371/journal.pone.0181001, (2017).

[18] S. Al-Shamsi, D. Regmi, and R. D. Govender, ''Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: A population-based study,'' PLoS ONE, vol. 13, no. 6, Jun. 2018, Art. no. e0199920.

[19] Z. Yang, J. Ren, Z. Zhang, Y. Sun, CH. Zhang, M. Wang, and L. Wang. A New Three-Way Incremental Naive Bayes Classifier. Electronics. 12. 1730. 10.3390/electronics12071730. (2023)

[20] T. Dai and Y. Dong, "Introduction of SVM Related Theory and Its Application Research," 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Shenzhen, China, pp. 230-233, doi: 10.1109/ AEMCSE50948.2020.00056. 2020

[21] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery", Springer, 2(2), pp.121-167, 1998.

[22] M. A. Abdelaal, M. A. Fattah, and M. M. Arafa. "Predicting Sarcasm and Polarity In Arabic Text Automatically: Supervised Machine Learning Approach." Journal of Theoretical and Applied Information Technology 100.8 (2022).

[23] Garavand, Ali & Behmanesh, Ali & Aslani, Nasim & Sadeghsalehi, Hamidreza & Ghadezadeh, Mustafa. (2023). Towards Diagnostic Aided Systems in Coronary Artery Disease Detection: A Comprehensive Multiview Survey of the State of the Art. International Journal of Intelligent Systems. 1-19. 10.1155/2023/6442756. 2023

[24] A.G. Karegowda, V. Punya, M.A. Jayaram, A.S .Manjunath," Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", International Journal of Computer Applications (0975 – 8887) Volume 45– No.12, May 2012.

[25] A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, AH. Zadeh, M. Ghaderzadeh. "Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms". J Healthc Eng. Oct 18;2022:5359540. doi: 10.1155/2022/5359540. PMID: 36304749; PMCID: PMC9596250. 2022.

[26] Maalouf, Maher. "Logistic regression in data analysis: An overview". International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335. 2011.

[27] T. Soliman A. Abd-elaziem. "A Multi-Layer Perceptron (MLP) Neural Networks for Stellar Classification: A Review of Methods and Results". International Journal of Advances in Applied Computational Intelligence. 3. 10.54216/IJAACI.030203. 2023.

[28] D.S. Kumar, G. Sathyadevi, and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011 ISSN (Online): 1694-081.

## APPENDICES

### A. *Appendix A*

TABLE V. FREQUENCIES, PERCENTAGES, MEANS, AND STANDARD DEVIATIONS OF THE RESPONDENTS' RESPONSES TO TECHNICAL INFRASTRUCTURE QUESTIONS

| N. | phrase | Response | | | | Average | Standard deviation | Phrase arrangement | Response degree |
|---|---|---|---|---|---|---|---|---|---|
| | | | Disagree | Neutral | Agree | | | | |
| 22 | Kuwait is considered a high-income country that has advanced healthcare infrastructure | F | 3 | 6 | 21 | 2.60 | .675 | 5 | High |
| | | % | 10.0 | 20.0 | 70.0 | | | | |
| 23 | There is a need for the establishment of a well-designed primary health care network. | F | 0 | 4 | 26 | 2.87 | .346 | 2 | High |
| | | % | 0.0 | 13.3 | 86.7 | | | | |
| 24 | The use of computerized information systems in different healthcare facilities | F | 2 | 5 | 23 | 2.70 | .596 | 3 | High |
| | | % | 6.7 | 16.7 | 76.7 | | | | |
| 25 | Health informatics is already implemented in the State of Kuwait. | F | 3 | 11 | 16 | 2.43 | .679 | 7 | High |
| | | % | 10.0 | 36.7 | 53.3 | | | | |
| 26 | The existence of reliable registration, licensing, and authorization systems for medical information. | F | 3 | 10 | 17 | 2.47 | .681 | 6 | High |
| | | % | 10.0 | 33.3 | 56.7 | | | | |
| 27 | The need to employ e-health solutions to improve the quality of the healthcare sector. | F | 0 | 3 | 27 | 2.90 | .305 | 1 | High |
| | | % | 0.0 | 10.0 | 90.0 | | | | |
| 28 | The huge investment that is being allocated to develop the technical infrastructure of the State | F | 3 | 5 | 22 | 2.63 | .669 | 4 | High |
| | | % | 10.0 | 16.7 | 73.3 | | | | |
| Overall average | | | | | | 2.66 | .409 | -- | High |

*C. Appendix B*

TABLE VI. FREQUENCIES, PERCENTAGES, MEANS, AND STANDARD DEVIATIONS OF THE RESPONDENTS' RESPONSES TO ORGANIZATION AND ADMINISTRATION QUESTIONS

| N. | phrase | Response | | | | Average | Standard deviation | Phrase arrangement | Response degree |
|---|---|---|---|---|---|---|---|---|---|
| | | | Disagree | Neutral | Agree | | | | |
| 29 | Kuwait Vision 2035 pays great attention to the development of the national healthcare system. | F | 3 | 10 | 17 | 2.47 | .681 | 7 | High |
| | | % | 10.0 | 33.3 | 56.7 | | | | |
| 30 | The advancement in health research capacities in the State of Kuwait needs more interest in information technology | F | 1 | 3 | 26 | 2.83 | .461 | 3 | High |
| | | % | 3.3 | 10.0 | 86.7 | | | | |
| 31 | The need for high-quality services in healthcare public and private sectors. | F | 0 | 2 | 28 | 2.93 | .254 | 1 | High |
| | | % | 0.0 | 6.7 | 93.3 | | | | |
| 32 | The need for a national health information management strategy to guide public health policies. | F | 0 | 4 | 26 | 2.87 | .346 | 2 | High |
| | | % | 0.0 | 13.3 | 86.7 | | | | |
| 33 | The emergence of medical equipment manufacturers has appeared as a promising field in the State. | F | 4 | 7 | 19 | 2.50 | .731 | 6 | High |
| | | % | 13.3 | 23.3 | 63.3 | | | | |
| 34 | The healthcare system in Kuwait depends on an integrated information and technology framework. | F | 7 | 8 | 15 | 2.27 | .828 | 8 | High |
| | | % | 23.3 | 26.7 | 50.0 | | | | |
| 35 | The high priority that is given to the healthcare sector to improve the quality of services being provided to citizens | F | 4 | 6 | 20 | 2.53 | .730 | 5 | High |
| | | % | 13.3 | 20.0 | 66.7 | | | | |
| 36 | The huge investment that is being allocated to develop the technical infrastructure of the State. | F | 3 | 8 | 19 | 2.53 | .681 | 4 | High |
| | | % | 10.0 | 26.7 | 63.3 | | | | |
| Overall average | | | | | | 2.62 | .473 | -- | High |

*D. Appendix C*

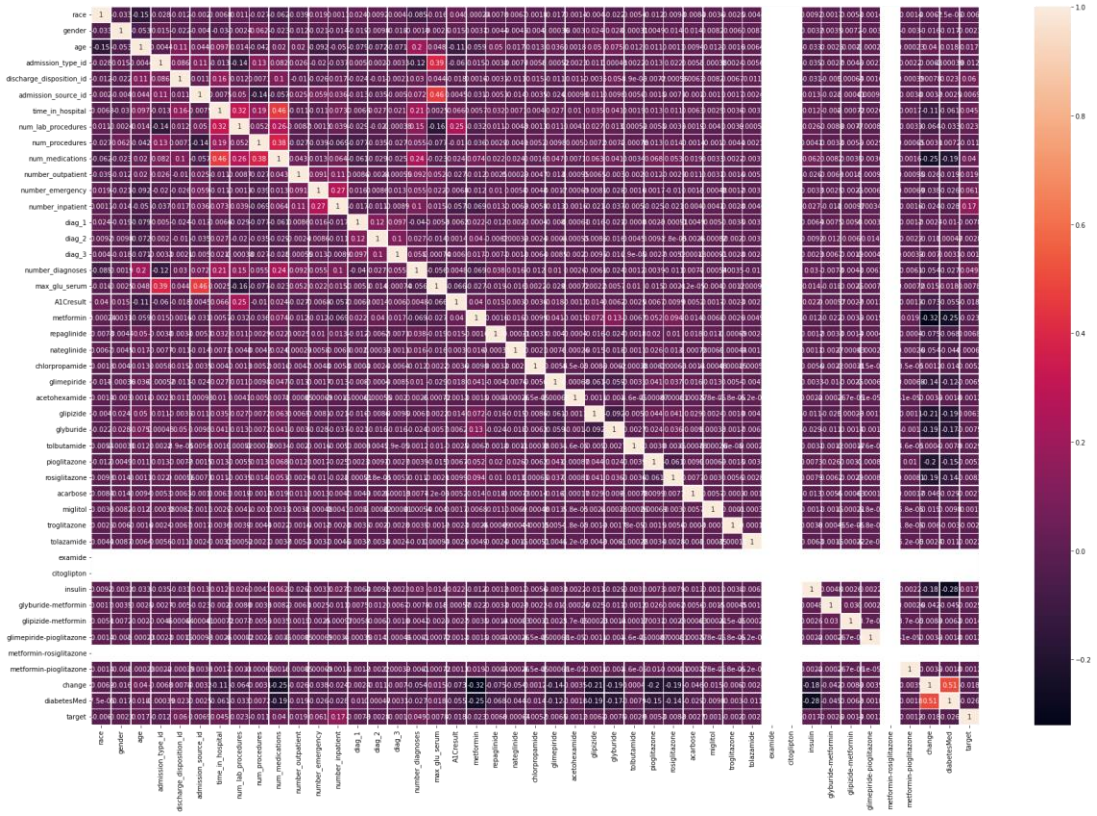


Fig. 4. Box Plot of features to see the outliers in diabetes dataset.