Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 5s (2024)

Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset

Merdin Shamal Salih[1], Rowaida Khalil Ibrahim[2 *], Subhi R. M. Zeebaree[3], Dilovan Asaad Zebari[4], Lozan M. Abdulrahman[5], Nasiba Mahdi Abdulkareem[6]

[1]Information Technology Dept., Technical College of Informatics-Akre, Akre University for Applied Science, Akre , 42004, Iraq, merdin.shamal@auas.edue.krd

[2]Computer Science Dept., Faculty of Science, University of Zakho, Zakho, 42002, Iraq, rowaida.ibrahim@uoz.edu.krd

[3]Energy Eng. Dept., Technical College of Engineering, Duhok Polytechnic University, Duhok, 42001, Iraq, subhi.rafeeq@dpu.edu.krd

[4]Computer Science Dept., College of Science, Nawroz University, Duhok, 42001, Iraq, dilovan.majeed@nawroz.edu.krd

[5]Information Technology Dept., Duhok Technical College, Duhok Polytechnic University, Duhok, 42001, Iraq, lozan.abdulrahman@dpu.edu.krd

[6]Information Technology Dept., Duhok Technical College, Duhok Polytechnic University, Duhok, 42001, Iraq, nasiba.mahdi@dpu.edu.krd

Corresponding Author: Rowaida Khalil Ibrahim, rowaida.ibrahim@uoz.edu.krd

Article History:

Received: *18-05-2024*

Revised: *16-06-2024*

Accepted: *02-07-2024*

Abstract

Diabetes mellitus, a chronic condition, causes disruptions in the metabolic processes of carbohydrates, lipids, and proteins. Hyperglycemia, characterised by elevated blood sugar levels, is the primary distinguishing characteristic of all forms of diabetes. Diabetes is a disease that has significantly increased in prevalence due to the contemporary lifestyle. Consequently, it is essential to get an early-stage diagnosis of the illness. When constructing classification models, data pre-processing is a crucial step. The Pima Indian Diabetes dataset, available in the University of California Irvine (UCI) repository, is a challenging dataset with a higher proportion of missing values (48%) compared to comparable datasets. To improve the accuracy of the classification model, many rounds of data pre-processing are conducted on the Pima Diabetes dataset. The proposed approach consists of two stages: outlier removal and imputation in the first stage, and normalisation in the second stage. Regarding the feature aspect, we used a method called principal component analysis (PCA). Ultimately, to classify the PIMA dataset, we used many classifiers such as Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and Decision Tree (DT). The testing revealed that the maximum achievable accuracy was 89.86% when 80% of the data was used for training. This was accomplished by integrating the feature selection technique with the classifier.

Keywords: Diabetic, PIMA dataset, Data pre-processing, PCA, PIMA dataset, Classification.

1. Introduction

1.1. Diseases Prediction Based on Machine Learning

The use of machine learning (ML) in illness prediction has fundamentally transformed the healthcare business. Machine learning algorithms can use extensive medical data to detect patterns and produce

accurate predictions on the beginning, course, and outcomes of diseases. This theoretical review explores the core principles, approaches, and difficulties associated with using machine learning for illness prediction. Machine learning is a subdivision of artificial intelligence that focuses on constructing systems that can acquire knowledge from data and make forecasts or choices without requiring explicit programming. Important principles in machine learning encompass: The mathematical methodologies used to represent the data. Popular algorithms used in illness prediction encompass logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. The outputs of the learning process are the patterns identified in the training data. Training entails acquiring knowledge from a dataset, while testing assesses the model's effectiveness on unfamiliar data. The quantifiable attributes or traits of the observed phenomenon. When predicting diseases, the elements that are taken into consideration may consist of patient demographics, medical history, genetic information, and lifestyle variables. The model's objective is to forecast the outcomes or categories. In the context of illness prediction, labels often indicate whether a certain disease is present or not[1].

Machine learning approaches used in illness prediction may be roughly classified into three categories: Models are trained using labelled data, which consists of known input characteristics and their matching output labels. The most prevalent method in illness prediction involves activities such as categorising people as either sick or healthy. Models use unsupervised learning to detect patterns in data without labelled outputs. This methodology is valuable for uncovering hidden patterns in data, such as grouping patients with comparable symptoms. Models acquire decision-making abilities via their interaction with an environment and the subsequent input they get. Although less prevalent in illness prediction, it may be advantageous for tailoring treatment regimens to individual needs[2].

Various strategies and techniques are used in illness prediction via the application of machine learning: The first phase is cleansing and preparing the data for analysis. This encompasses the tasks of managing null values, standardising data, and converting categorical variables into numerical representations. The identification and creation of relevant characteristics are essential for optimising model performance. Methods such as principal component analysis (PCA) and recursive feature elimination (RFE) are used to identify significant features. The selection of the most suitable machine learning method is contingent upon the characteristics of the task at hand and the available data. Logistic regression is appropriate for binary classification, but neural networks are more effective for intricate patterns. Models undergo training using a portion of the data and are then assessed for performance using a separate portion of the data. Methods such as cross-validation aid in ensuring that the model exhibits good generalisation to unfamiliar data. Standard measures used to evaluate model performance include accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC)[3].

Various machine learning techniques are often used for illness prediction: A logistic regression model is a statistical model that estimates the likelihood of a binary outcome by considering one or more predictor variables. It is straightforward and easily understood, which makes it valuable for developing early illness prediction models. Decision tree models partition the data into branches according to feature values in order to generate predictions. They are easily understandable yet susceptible to overfitting. A technique called ensemble learning that enhances accuracy and resilience by combining

numerous decision trees [4]. It mitigates overfitting in comparison to individual decision trees. An effective classification technique that identifies the most suitable hyperplane for separating distinct classes inside the feature space. Neural models that draw inspiration from the human brain and possess the ability to acquire intricate patterns. Deep learning, a specific branch of neural networks that consists of numerous layers, has shown significant potential in the field of illness prediction, particularly when applied to picture and genomic data [5].

Machine learning algorithms have been used to forecast a diverse array of illnesses, encompassing: Utilising patient data such as age, blood pressure, cholesterol levels, and lifestyle variables to forecast occurrences of heart attacks and strokes. Utilising medical imaging and genomic data for the prompt identification and categorization of tumours. Machine learning may be used to detect tumours in medical imaging and accurately forecast their aggressiveness. Anticipating the initiation of diabetes by the examination of patient data and lifestyle variables. Models have the ability to identify persons who are at a high risk and propose actions to avert negative outcomes. Identifying disorders such as Alzheimer's and Parkinson's by the examination of brain scans, genetic data, and cognitive test outcomes. Forecasting the dissemination of illnesses such as influenza and COVID-19 by the use of transmission patterns modelling and the examination of epidemiological data[6].

In order to properly use machine learning for illness prediction, it is necessary to solve many problems despite the promise it holds. Training robust models requires the use of high-quality, representative, and large-scale datasets. Nevertheless, medical data often lacks information, contains errors, or exhibits prejudice. A significant number of machine learning models, particularly those based on deep learning, function as opaque entities, hence posing challenges in comprehending the underlying mechanisms behind their predictions. Interpretability is essential for the acceptability of clinical applications [7]. Models must have strong generalisation capabilities across diverse populations and environments. Overfitting to individual datasets might restrict the generalizability of the models. Managing sensitive medical information requires rigorous privacy protocols and ethical deliberations to safeguard patient confidentiality and get permission. Integrating machine learning models into healthcare settings necessitates smooth incorporation into current systems and processes, as well as providing training for healthcare personnel [8].

Machine learning has great promise for illness prediction, since it can analyse intricate medical data and make early and precise predictions. Machine learning has the potential to greatly improve healthcare results and provide more individualised and efficient treatments by tackling present difficulties and using forthcoming developments.

1.2. Diabetic Prediction Based on Machine Learning

Since the beginning of time, health has been and will continue to be a top priority. Given the significant amount of progress that has been made in the healthcare industry, there is a great deal of room for research. It is imperative that the current healthcare technology be upgraded by embracing the digitalization of medical information. This applies to the data that is provided by patients as well as the medical outcomes that are generated by modern equipment. The fact that we are confronted with the onerous task of evaluating and comprehending the massive amounts of data that have been obtained

is a common consequence of this digital revolution. In light of the fact that there is an enormous quantity of data, Big Data Analytics appears to be of great assistance [9].

Diabetes mellitus (DB) is a debilitating health condition that places a large additional financial burden on healthcare providers all over the world in terms of the expense of treatment. Diabetes type 1, sometimes referred to as hyperglycemia, is a condition that manifests itself when the beta cells of the pancreas fail to secrete an adequate amount of insulin, resulting in elevated levels of glucose in the blood [10]. People who have diabetes type 2 have a body that is unable to make efficient use of the insulin that is available in their bodies. Furthermore, diabetic retinopathy may result in numerous clinical ramifications, such as harm to the neurological system, degeneration of the retina, renal sickness, and cardiovascular disease [11]. In 1980, there were 108 million people who were diagnosed with diabetes. In 2014, the number of persons afflicted by diabetes is estimated to have surpassed 422 million, which is a significant increase from that amount. In addition, over the same time period, the percentage of people who were diagnosed with diabetes (also known as people with diabetes) increased from 4.7% to 8.5% of the total adult population. This is a significant challenge for those who are attempting to control diabetes. In the year 2012, high blood glucose levels were the cause of death for 2.2 million persons who had diabetes [12]. One million and six hundred thousand people died as a result of diabetes in the year 2015. When it comes to achieving the objectives of maximizing treatment choices, increasing the quality of life for individuals who have diabetes, and decreasing death rates connected with the condition, timely diagnosis and early identification of diabetes are very essential. Diabetes is expected to become the sixth largest cause of death throughout the globe by the year 2030, according to projections. Furthermore, a sizeable proportion of people who have impairments do not become aware of their disease until a critical complication manifests itself; there is a correlation between the delay in identifying the start of type 2 diabetes in its early stages and an increased chance of serious outcomes [13].

A reliable model that is able to effectively depict the presence of diabetes through the features that are input is required in order to make an accurate prediction of the illness. It is possible to improve the efficiency of diagnosis by the utilization of a reliable model and a detection method that is precise. Using the forecast, medical professionals are able to foresee the possibility of doing biomedical diagnosis with the assistance of engineering tools that are able to automatically adjust to any unanticipated future situations. With regard to planning and provisioning, a long-term prediction algorithm has the potential to be of great assistance. In reaction to new experiences or shifts in the functional relationships between components, intelligence systems have the ability to learn, adapt, and adjust the functional dependencies, respectively [10, 13].

When it comes to early diagnosis, the forecasting and identification of the disease are assessed through the knowledge and expertise of a physician; however, this method is not without its limitations and can be subject to error. The healthcare industry gathers a massive amount of data pertaining to healthcare; however, this data is unable to recognize patterns that have not yet been discovered, which prevents it from making successful judgments [14]. Because decisions made manually are based on the observations and judgment of the healthcare official, which is not always accurate, manual decisions can be extremely risky when it comes to the early diagnosis of health conditions. There may be some patterns that are not readily apparent, which may have an effect on the observations and the

results. Patients are receiving a low quality of service as a consequence of this; hence, an advanced mechanism is necessary for early identification of illness with an automatic diagnosis and improved accuracy. In the process of data mining and machine learning, many flaws and hidden patterns that have not been discovered give rise to a wide variety of algorithms that are capable of producing efficient results with trustworthy accuracy [15]. A wide range of data mining techniques have been developed in response to the ever-increasing impact of diabetes on a daily basis. These algorithms are designed to uncover hidden patterns within enormous amounts of healthcare data. In addition, the data can be utilized for the purpose of selecting features and making automated predictions regarding diabetes. This research work's primary objective is to suggest an invention of a prognostic tool for early diabetes forecasting and diagnosis that is more accurate than previous methods. The PIMA dataset, which has been utilized in this work, is one of the most extensively used datasets. There has been a substantial amount of data and datasets that have been made available on the internet or from other sources. SVM, NB, DT, and RF are some of the data mining techniques that were utilized in this research effort, which represents thorough studies that were conducted on the PIMA datasets.

2. Literature Review

Data mining can be applied across various areas like health care, education, business, and numerous other domains. Data mining in healthcare facilitates the identification of illnesses, prognosis, and comprehensive analysis of medical data. For example, it can enhance comprehension of the relationship between several chronic illnesses, including type 2 diabetes (DM), which is a significant health issue and a leading cause of mortality.

The study [16] included a complete analysis of three different data mining algorithms that were used for the purpose of predicting diabetes or prediabetes. There were three types of models that were used in the data mining process: decision trees (DT), artificial neural networks (ANNs), and logistic regression (LR). There are 735 patients and 752 normal controls included in the dataset that was utilized. The distribution of the patients is even. Twelve parameters were used in the construction of the models. These parameters included gender, age, marital status, educational attainment, familial predisposition to diabetes, body mass index (BMI), coffee consumption, amount of physical activity, duration of sleep, stress related to work, fish consumption, and preference for salty foods. In order to collect the aforementioned characteristics, a survey was developed and administered. The findings of the study led the researchers to the conclusion that the C5.0 decision tree exhibited superior performance in terms of identification precision. Using the PIDD dataset, Abdulhadi et al. [17] used a number of different machine learning models in order to make predictions about the development of diabetes in females. The issue of missing data was handled by the authors via the use of the mean replacement strategy, and all of the attributes were standardized through the utilization of a standardization procedure. Logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM) with linear and polynomial kernels, and random forest (RF) were the methods that were used in the construction of the models. According to what is shown in the paper, the RF model achieved a peak accuracy score of 82%.

In order to accomplish accurate categorization of diabetes, N. K. Putri, Z. Rustam, and D. Sarwinda used the Learning Vector Quantization and Chi-Square feature selection techniques. By using between 80 and 90 percent of the training data, they were able to achieve a rate of accuracy of one hundred

percent [18]. Regarding the classification of cancer, T. Nadira and Z. Rustam used Support Vector Machines (SVM) in conjunction with feature selection approaches. As can be seen in the dataset [19], the accuracy rates that were achieved for lung cancer and breast cancer were respectively 99.9999% and 96.4286%. A classification accuracy of 90% was achieved by Arfiani, Z Rustam, J Pandelaki, and A Siahaan via the use of Support Vector Machines (SVM) in the categorization of acute sinusitis [20]. Support Vector Machines (SVM) were used by Rustam and Rampisela T. V. in order to categorize data pertaining to schizophrenia, and they achieved an accuracy rate rate of 90.1% [21]. HbA1c regression models were established in [22], which may be found here. The HbA1c test is a measurement that is used to determine the average concentration of glucose in the blood over a period of two to three months, as stated by the research. There is a substantial correlation between it and diabetes, and it may also act as a warning of potential future complications. The information that was used in this investigation was obtained from the Diabetes Research in Children Network (DirecNet) studies. These trials comprised 170 patients who were diagnosed with type 1 diabetes mellitus and were classified as being between the ages of 4 and under 10 years old. In order to solve the issue of missing data, the mean substitution method was used as a solution. In addition, any property that was lacking data by more than twenty percent was eliminated. Additionally, the dataset was subjected to a number of different approaches for the extraction and selection of relevant features. A low mean absolute error (MAE) of 3.39 mmol/mol and a high coefficient of determination (R-squared) score of 0.81 were attained by the ultimate machine learning (ML) model, according to the findings of the research. This model used two ensemble approaches, namely Random Forest (RF) and extreme gradient boosting (XGB).

Support Vector Machines (SVM) with many kernel functions were used by G.A. Pethunachiyar in order to classify diseases related to diabetes. The simulation model of the system that has been proposed is comprised of five phases. When the data has been collected, the next step in the selection process is to correct any faults that may have been present, such as inconsistencies in the data, numbers that are missing, or information that is erroneous. The data will be separated into two pools: a training dataset, which will consist of seventy percent of the data, and a testing dataset, which will consist of thirty percent of the data. The technique known as Support Vector Machine (SVM) has been chosen because of its capacity to produce accurate forecasts, and a model has been constructed in accordance with this particular approach. The predictions are generated by the model based on the test data. In this study, Support Vector Machines (SVM) have been used, and linear, polynomial, and radial kernel functions have been utilized. For the purpose of determining how accurate predictions are, the confusion matrix is used. In order to analyze three different kernel functions, the ROC curve is used. When compared to other kernels, the linear kernel used in Support Vector Machines (SVM) offers a higher level of accuracy within its prediction capabilities [23].

One of the most important applications of machine learning is in the diagnosis and prediction of potential diseases. Through the development of an Artificial Neural Network (ANN) and a Bayesian network, Alade et al. [24] presented a method for predicting the occurrence of diabetes. For the purpose of training and evaluating the dataset, the artificial neural network (ANN) architecture is comprised of four layers and makes use of the back-propagation method and the Bayesian regulation algorithm. Careful training has been performed on the data in order to ensure that the conclusions on the

regression graph are appropriately represented. The use of this technology enables remote diagnosis, which in turn makes it easier to communicate with patients without the need of being physically near to them. The research that was carried out by Stefan Ravizza and his colleagues [25] investigates the use of data mining techniques in the field of healthcare and suggests a model for assessing the possible dangers of illnesses that are not appropriately treated. The approach that they used in the healthcare industry is one that is based on empirical evidence, is directed by certain characteristics, and requires the collecting of data from live situations. Within the context of the Deep Patient approach, this method is distinct. The findings of this strategy were compared with clinical data utilizing direct algorithms, and the results were discussed.

The PIMA dataset and advanced machine learning techniques were used in this research project with the intention of accurately predicting the occurrence of problems related to diabetes in individuals who experience the condition. The impacts of applying feature selection to the dataset were investigated via a series of experiments that were carried out in order to assess various data imputation methods, balancing procedures, and effects.

3. Proposed Method

A condensed review of the advancements that have been made in the use of technology is provided in this section. Individuals who have diabetes are the primary recipients of notifications from the proposed classifier model, which also incorporates input into the diabetes dataset. In the first step of this process, we gather the contextualized dataset of Pima Indian diabetes. For the purpose of gaining a full knowledge of the sources from which our data comes, exploratory data analysis is carried out. The following step is to preprocess our data by purging our dataset, specifically by eliminating any duplicate, missing, or anomalous values that could be present. This is a critical step in the process. After that, we will choose the models that will be used to train our data, and then we will put the model that we have selected into action. Following this, the models will be evaluated and compared based on a variety of performance indicators, such as accuracy, sensitivity, and specificity amongst the models. The strategy that is proposed is shown in Figure 1, which provides an overview of the stages that are engaged in the implementation process step by step.

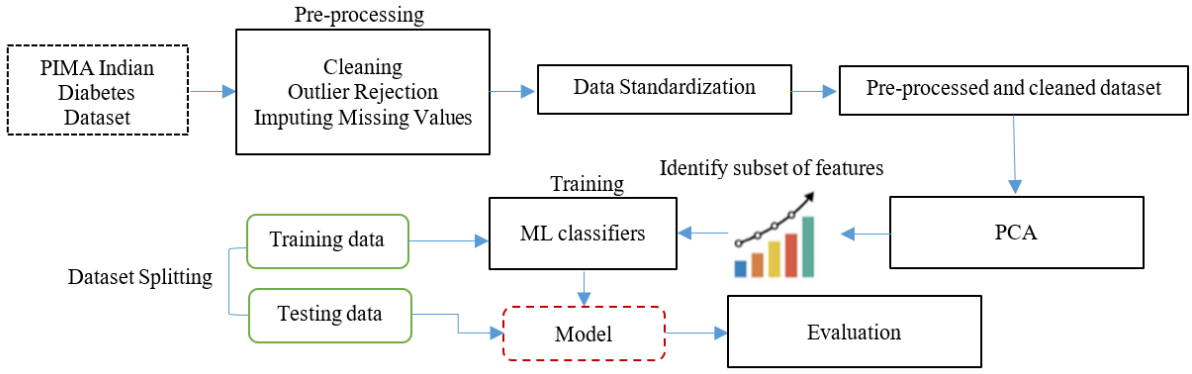

Figure 1: Overview of the Proposed Model

In the beginning, the data was cleaned up by deleting instances that included a substantial quantity of values that were missing. Following this, mean/mode imputation was used in order to effectively replace the data that was absent. With regard to the process of imputing missing data, this technique has shown excellent outcomes. Through the use of dummy variables, the categorical categories were transformed into numerical values. The variables were normalized to a consistent scale by the use of a feature scaling approach known as Standardization. This was done in order to address the discrepancies in the units and ranges of the variables. Following the process of normalizing all of the variables, the dimensionality reduction method known as Principal Component Analysis (PCA) was used. These variables were connected to a smaller set of independent principle components, and the purpose of this technique was to turn those variables into those principal components. The present issue might be characterized as a multilabel classification task, in which a patient may experience either one of the outcomes or both of them. Subsequently, classification techniques such Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest, and Decision Tree were used in order to forecast the probability of a patient's risk by using the PIMA dataset. A number of criteria, such as the Accuracy Score (ACC), the Confusion Matrix, the Classification Report, the ROC Curve, and the AUC score, were used in order to ensure that each algorithm was evaluated for its effectiveness.

3.1 Pre-processing

To begin processing the information, the initial step involves systematically deleting superfluous entries and characteristics using a cleaning approach [26]. Initially, the information contains several category characteristics that must be removed to ensure anonymity. The variables consist of the hospital number, the event date, and the occurrence description. Moreover, the dataset is deficient in data about the specific kind of diabetes in particular patients, a crucial piece of information for our study since we examined difficulties specifically related to diabetes in individuals with the condition. Consequently, all twenty-six occurrences that were impacted by this issue were eliminated. Subsequently, a Z-score, also known as a standard score, is used to standardize scores on a uniform scale. To get this measurement, the variance of a score is divided by the standard deviation of a data collection. The term "standard deviation" is used to measure the extent to which a certain data point diverges from the mean. This variance is expressed in terms of the standard deviation. A negative Z-score signifies values that are below the mean, a positive Z-score signifies values that are above the mean, and a Z-score of zero signifies the mean value. The value is within the range of (-1, 1), with a positive Z-score indicating values above the mean. Z-score normalization may be calculated using the mean (μ) and standard deviation (σ) of the characteristics. During the calculation, the feature vector x is used as the beginning value [27].

3.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a technique that may be used to decrease the number of variables in a dataset. It does this by applying an orthogonal transformation, resulting in a reduced set of variables while retaining much of the original information. The procedure entails reducing the quantity of highly linked variables in the original dataset to a reduced number of linear variables that are unconnected to each other. These variables are often known as principal components. Subsequently, these primary constituents are accountable for the bulk of the variability seen in the original dataset [28, 29]. Principal component analysis (PCA) is particularly advantageous when

dealing with data that has three or more dimensions, as it becomes increasingly challenging to make predictions based on such a substantial amount of information. Moreover, presenting data that has several dimensions is a significant challenge. Furthermore, PCA has the capability to address this problem in terms of data visualization. The number of primary components is determined by the lower value between the number of observations and the number of original features. Initially, the dataset consisted of 779 observations and 164 relevant features. Consequently, the maximum number of principal components that could be used was 164. Out of the 164 components, a subset of fifty was chosen. Figure 2 clearly shows that these fifty components contributed to eighty percent of the overall variance.

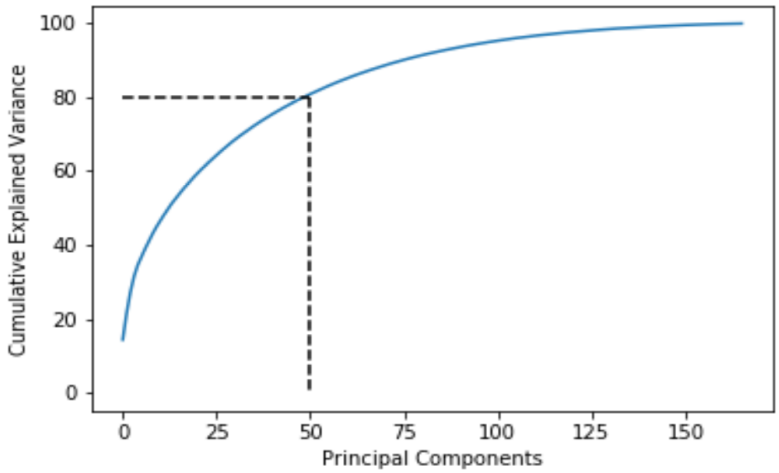

Figure 2. Cumulative Explained Variance against No. of Principal Components [30].

3.3 Machine Learning

Supervised learning is a versatile technique that may be used to any kind of data without limitations. Classification is a sequential procedure that first acquires information from the provided data and then applies that knowledge to categorize new data. This technique is beneficial for selecting the appropriate class labels for integrating new data. The clustering methodology enables the creation of diverse labels and groupings within the dataset, relying on the existing commonalities among the data points. Classification is a technique used in data mining to categorize data, enabling more precise analysis and predictions. Machine learning [31] is a data mining approach specifically designed to analyze very large datasets. To accurately classify the important data categories in the data collection, patterns are generated. Classification algorithms provide predictions for the target classes of every occurrence in the given dataset. Classification algorithms aim to examine data and establish connections between traits to enable precise prediction of outcomes. An analysis is conducted on the input, which leads to the generation of a forecast. Data mining activities including classification are often used in the healthcare sector [32].

3.3.1 Support Vector Machine

A support vector machine (SVM) is a linear model that may be used to address classification and regression problems. With this tool, users may discover answers to both linear and nonlinear problems.

Similar to linear regression, it functions in an analogous fashion [33]. The SVM approach classifies new data by generating a hyperplane with the largest possible margin.

The main goal of these supervised algorithms is to create an unparalleled decision boundary or line that can divide n-dimensional space into distinct classes. In the future, we will be able to easily assign the new data instance to the correct group. A hyperplane refers to the decision boundary that cannot be surpassed. Multiple alternative boundaries may be designed, as seen in Figure 3, to separate the classes in n-dimensional space. However, the primary objective should be to identify the ideal decision boundary that enables the most efficient categorization of the data points.

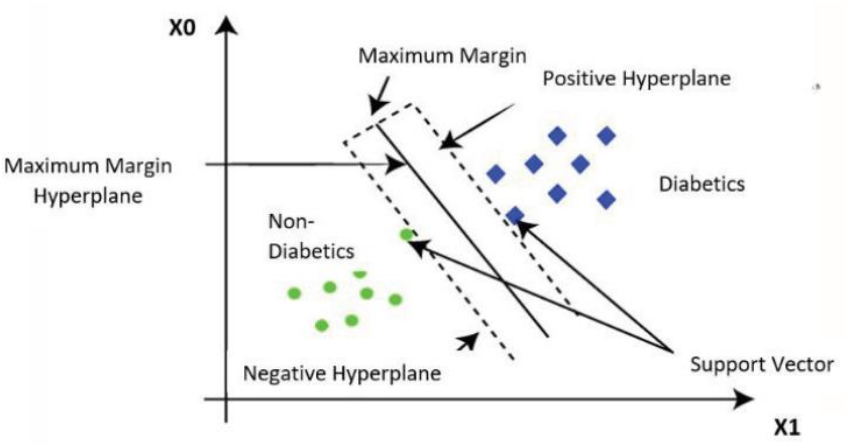

Figure 3: Illustrating the SVM Hyperplanes [34]

Support vector machines (SVM) do classification problems by creating hyperplanes in a multidimensional space that distinguish between occurrences of distinct class labels. SVM refers to the set of data instances that provide the foundation for the hyperplane. The word "margin" denotes the spatial separation between the hyperplane and the support vector that is in closest proximity to it. When selecting the optimal hyperplane, the primary factor considered is the maximum margin, which represents the largest separation distance between the two classes. Both linear and non-linear SVMs are used to classify data points. Linear SVMs are used to classify data points by using a straight decision boundary, whereas non-linear SVMs employ a curved decision boundary to classify data points [35, 36].

The data points (X1, Y1)....(Xn,Yn) consist of real vectors represented by xi and binary values represented by y1, where y1 takes the value of either 0 or 1. The value of y1 indicates the class to which xi belongs. The building of a hyperplane is to maximize the distance between two classes, y = 0 and 1. Its definition is as follows:

$$max_{\cup a} \sum_{i=1}^{n} \square\, a_i - \frac{1}{2} \sum_{i=1}^{n} \square \sum_{j=1}^{n} \square\, y_i y_j K(x_i, x_j) a_i a_j \qquad (1)$$

Subject to:

$$0 \le a_i \le c,\ for\ i = 1,2, \ldots n, \sum_{i=1}^{n} \square\, y_i a_i = 0 \qquad (2)$$

3.3.2 Naïve Bayes (NB)

The Naïve Bayes method is a supervised learning approach that utilizes Bayes' Theorem for classification purposes. Bayes' theorem utilizes conditional probability, which relies on previous knowledge, to assess the chance of a future event occurring. Bayes' Theorem may be expressed using the following formula:

$$P(E) = \frac{P(E)*P(H)}{P(E)}$$

In this application, the posterior probability, represented as P(H|E), refers to the likelihood that a hypothesis (H) is accurate given specific evidence (E). The prior probability, sometimes referred to as the probability that the hypothesis is true, is represented by the sign P(H). The sign P(E) represents the probability of the evidence, regardless of the hypothesis. The conditional probability of the evidence given that the hypothesis is true is represented by the notation P(E|H) [35].

The Naïve Bayes classifier is a classification approach that assumes the input variables (features) are independent of each other and that each feature contributes individually to the probability of the target variable. It may be inferred that the existence of a solitary feature variable does not have any influence on the other feature variables. As a result, it is often referred to as naive. Due to the interdependence of feature variables in real datasets, the Naïve Bayes classifier now faces this specific constraint. The Naïve Bayes classifier has exceptional performance when used to large data sets and may even surpass more intricate classifiers in some scenarios. The Gaussian Naïve Bayes classifier, a specific kind of Naïve Bayes classifier, was used in this model. The Gaussian Naïve Bayes classifier assumes that the feature values are continuous and that the values pertaining to each class follow a normal distribution [36, 37].

An important characteristic of the Naïve Bayes approach is its ability to be trained on a little dataset, which is a notable benefit. This methodology is used for both binary and multiclass classification tasks. Furthermore, it is very efficient and easily adaptable to larger scales. Furthermore, it aids in mitigating the challenges arising from the curse of dimensionality to some degree. However, based on the previous statement, it assumes without sufficient evidence that the input variables are unrelated to one other. In contrast, real-world datasets may include several complex interactions among the feature variables. However, some datasets do not follow this pattern.

3.3.3 Random Forest

An ensemble of models is the most important aspect of RF; each model makes a prediction about the result, and in the end, the model that has the majority of its predictions true is the winner. RF is a collection of models that may be performed together in a manner similar to that of an orchestra [33]. During the course of the study that has been conducted on the topic, it has been investigated, and it has been proven to be useful in predicting diabetes. There is a learning model known as a Random Forest that is both flexible and capable of solving issues involving regression and classification. When Random Forest is in the training phase, it first creates a huge number of DT and then it makes a prediction that is an average of all of the forecasts that were made by the decision trees. Both types of difficulties are within the capability of this model [37], which means that they are applicable. When dealing with classification issues, the goal variable is categorical, but when dealing with regression, it

is continuous. During regression, continuous variables are used. EDA methods are applied in order to accomplish the goal of achieving a high degree of accuracy via the use of Random Forest. A more robust model is produced as a consequence of the combination of a large number of models that are relatively weak. The method that is often referred to as "Random Forest" is notable for its ability to analyze big datasets that include a significant amount of geographical information. The dimensionality reduction capability of the model is the capacity of the model to analyze a large number of input variables and identify which ones are the most important. This capability includes the ability to decide which variables are the most significant. It is a valuable feature of the model because it shows the significance of variables even when dealing with a random dataset. This is one of the model's characteristics. In contrast to the traditional model, which employs a voting strategy that is equivalent to one another, the "Adaptive Random Forest" (ARF) model is the one that accomplishes the greatest results when an uneven voting strategy is used [38].

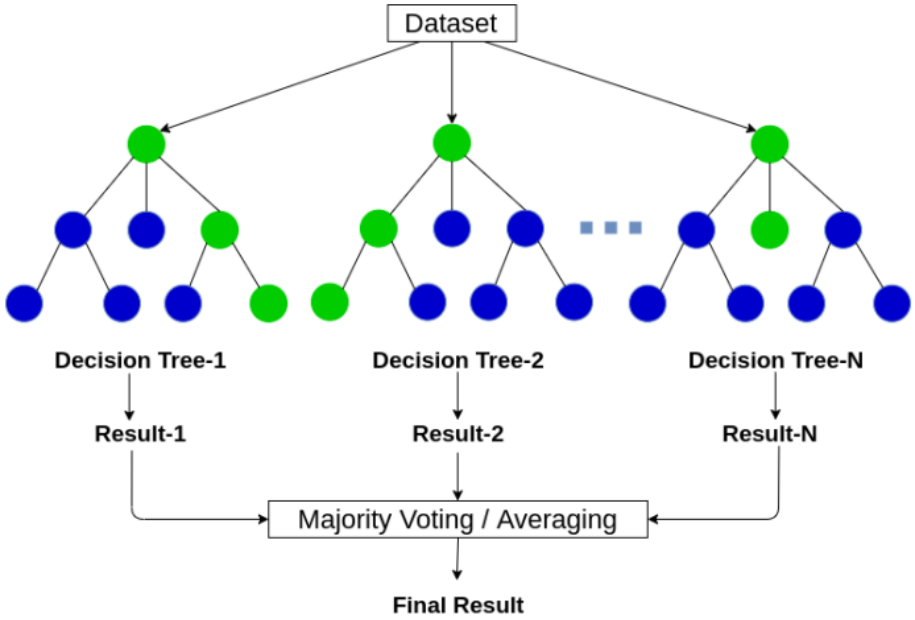

Figure 4: Random Forest structure [38].

3.3.4 Decision Tree

A decision tree method, often known as a DT classifier, is a kind of decision-making assistance. This approach is implemented using a hierarchical structure resembling a tree, and it is built by including certain input properties [36]. The primary goal of this classifier is to construct a model that can make predictions about the targeted variables based on a range of input attributes. This classifier is suitable for a wide variety of applications [37]. The reason for this is because it is quite simple to construct decision rules based on a given set of input data. The DT methodology is a nonparametric supervised learning method that may be used to address difficulties like as regression and classification, when implemented correctly. Figure 4 presents a representation of a structure, demonstrating the possible interpretation of the DT model. This specific paradigm consists of three nodes: the root node, the division node, and the leaf node. Each internal node functions as a test that is conducted on a certain

attribute. This test yields the outcomes for every division, and each individual leaf node retains the class label of its parent.

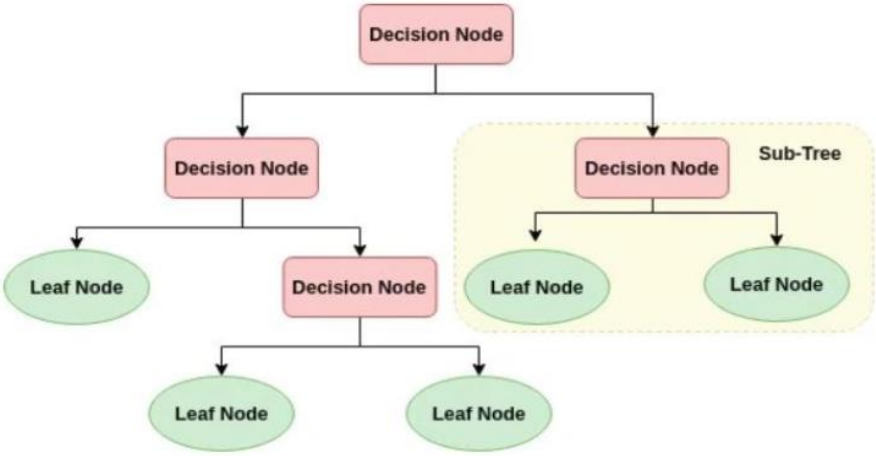

Figure 5: Decision Tree Structure [39].

Each division corresponds to the outcome of that exam. The root node marks the beginning of the tree construction process. Initially, an attribute is selected to be positioned at the root node, where it will persist during the whole method as the initial point. After completing the division, each of the potential values is submitted to it. As a result, the dataset is divided into subgroups, with each subgroup representing one of the several possible values found in the property. The tree operation is executed recursively for each division, considering only instances that have reached the branch. When all instances on a node are categorized uniformly, it becomes straightforward to halt the progress of the tree. When attempting to determine the optimum tree partition, it is customary to utilize either entropy or classification error as the preferred metrics [33].

3.4 Model Training

During the model training phase of our technique, which is centered on the construction of predictive models for diabetes diagnosis, we make significant progress. Additionally, in order to construct reliable prediction models, we have integrated a number of different machine learning algorithms, as was mentioned before. The method starts with the importing of the diabetic dataset and any relevant libraries required for the process. After this, the stages of data preparation are carried out, which include the elimination of any missing data and the normalization of the data in order to guarantee the quality and consistency of the data. In addition, we carry out two percentage splits: the first separates the dataset into forty percent for training and twenty percent for testing, and the second splits it into seventy percent for training and thirty percent for testing. Principal Component Analysis (PCA) is then used for the purpose of feature selection, which ultimately results in an increase in the effectiveness of our models. As we go further, a variety of machine learning algorithms, such as Support Vector Machine, Decision Tree, Random Forest, and Naïve Bayes (NB), are chosen for the purpose of developing models. Following the construction of classifier models for each method based on the training set, the models are next subjected to rigorous testing on the test set in order to evaluate their performance. With the purpose of determining the effectiveness of each classifier, a comparative analysis of the outcomes of the experiments is carried out. Finally, after doing an in-depth investigation

based on a variety of performance measures, we have arrived at the conclusion that the algorithm that performs the best for diabetes prediction is. The use of this all-encompassing methodology guarantees the selection of the most suitable model for the precise and dependable identification of diabetes.

4. Experimental Result

The results of the experiments conducted on the suggested models are presented in this section. In order to construct the prediction models, the SVM, RF, DT, and NB algorithms were applied. The performance of these models was assessed by using the appropriate metrics, which included accuracy, precision, recall, and F1-score. For the purposes of training and testing, we used a splitting data distribution, and we utilized two different splitting ratios. There is a 70% training and 30% testing ratio, and 80% training and 20% testing. By training and testing a model, one may increase the likelihood of attaining a successful outcome with the prediction. For the purpose of training algorithms that are capable of learning and making predictions, the training dataset is a collection of learning sets that are necessary. It is primarily for the purpose of evaluating the effectiveness of the chosen classifiers that the test set is used. The sole reason it is included is for the purpose of testing the classifiers. The accuracy that is anticipated to be achieved by a model is improved if it performs better in both datasets.

4.1 Dataset

The dataset resulting from the Pima Indian diabetes study is named "Diabetes.csv". This dataset was generated from diabetes. Diabetes may be recognized by its eight distinctive characteristics, which function as indicators. These findings are derived from a dataset of 768 cases and are used to ascertain the presence or absence of diabetes in patients. Regarding patients, this norm might manifest in several ways. The table below displays the indicators of the physical features of the data set. Table I is a succinct explanation of the eight attributes that constitute the diabetes dataset [14]. The research investigation has confirmed that the diabetic database file, which was helped by pandas, has been accessed. The database consists of 768 entries, each containing eight medical prediction parameters as input, and one target variable output. The target variable represents the presence of diabetes, with a value of one indicating "yes" and zero indicating "no". It has been noted that among a group of 768 Pima Indian women, 65.1% of them had not received a diagnosis of diabetes. The data is shown in figure 4, illustrating this discovery. On the other hand, a total of 34.90 percent of the 768 Pima Indian women were found to have been diagnosed with diabetes [29].

Table 1: Attribute of Dataset

Attribute	Description	Type	Average/Mean
Preg	Number of times pregnant.	Numeric	3.85
Glucose	Plasma glucose concentration 2 h in an oral glucose tolerance test.	Numeric	120.89
BP	Diastolic blood pressure (mm Hg).	Numeric	69.11
SkinThickness	Triceps skinfold thickness (mm).	Numeric	20.54
Insulin	2-hour serum insulin (μIU/mL).	Numeric	79.80
BMI	Body mass index (kg/m^2).	Numeric	32
DPF	Diabetes pedigree function.	Numeric	0.47
Age	Age (years).	Numeric	33
Outcome	Diabetes diagnose results (tested_positive: 1, tested_negative: 0)	Nominal	–

4.2 Results and Discussion

It is possible to examine a variety of performance indicators, such as accuracy, sensitivity, specificity, and error rate. Accuracy may be defined as the correlation between the total number of predictions and the fraction of correct forecasts. It is possible to define sensitivity as the percentage of positive samples that are found to be positive after testing. The term "true positive rate" is another name for this proportion. Specificity may be defined as the percentage of samples that are negative that are successfully tested negative. Additional names for this kind of rate include a real negative rate. It is possible to get the formula in Equation (3).

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (3)$$

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

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$F1-Score = \frac{2TP}{2TP+FP+FN} \quad (6)$$

The findings of the performance assessment of this research are shown in this part. The evaluation is based on the accuracy, sensitivity, and specificity of various matrices using the support vector machine (SVM) classifier. The assessment was carried out in two stages: first, before being submitted to PCA, and then again after being submitted to PCA. Following the construction of the PIAM diabetes dataset, the findings that were achieved for all classifiers are shown in the tables that are located below (Tables 2 and 3). According to the classification results that were obtained, Table 2 displays the results of 70% training and 30% testing. On the other hand, Table 3 displays the results of 80% training and 20% testing.

The training and testing split were determined to be thirty percent, regardless of whether PCA was used or not. The assessment method utilizes techniques such as SVM, NB, RF, and DT. Accuracy, precision, recall, F1-score, and AUC are measures used to assess the success of any procedure. The study of Table 2 reveals that Random Forest consistently achieves exceptional performance across all metrics, including accuracy, precision, recall, and AUC, independent of the use of PCA. SVM has exceptional performance, particularly when combined with PCA, resulting in outstanding accuracy

and AUC. When compared to other algorithms, Naive Bayes demonstrates a somewhat stable performance, while Decision Trees tend to exhibit poorer accuracy and AUC.

Table 2: All Model Performance for the 70% and 30% of training and testing ratio.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Without Using PCA					
SVM	76.32	72.53	77.65	70.91	85.49
NB	75.42	71.76	78.16	75.55	87.63
RF	82.42	85.33	74.62	77.38	79.32
DT	73.57	68.55	71.43	70.87	74.65
Without Using PCA					
SVM	85.54	87.33	86.47	89.12	91.13
NB	84.54	88.55	86.42	84.76	88.55
RF	88.76	87.33	89.13	90.22	92.43
DT	80.38	74.32	74.73	74.30	76.77

With an allocation of 80% for training and 20% for testing, Table 3 shows performance metrics that are equivalent to those shown in Table 2. The SVM, NB, RF, and DT algorithms are evaluated whether or not they make use of principal component analysis (PCA). Upon closer inspection, Table 3 exhibits patterns that are comparable to those seen in Table 2, which indicates that Random Forest consistently achieves good performance across all criteria. Performance that is constantly resilient is shown by the Support Vector Machine technique, especially when it is used in conjunction with Principal Component Analysis (PCA). Decision Trees have a little lower accuracy and area under the curve (AUC), but Naive Bayes continually maintains a high level of performance.

When the data produced from both tables are compared, it is evident that the Random Forest method constantly produces greater performance when compared to other algorithms. This is the case regardless of the training and testing ratios that are used. It is clear that lowering the number of dimensions is an effective method for enhancing classification abilities, as shown by the strong performance of the Support Vector Machine algorithm, particularly when Principal Component Analysis (PCA) is used. While Decision Trees have intermediate performance, they often have lower accuracy and area under the curve (AUC), Naive Bayes is reliable and reliable all the time, and its performance is constant. The purpose of these insights is to give useful guidance in selecting the best appropriate machine learning algorithm based on the specific aims and features of the dataset. In addition to taking into consideration the potential impact of PCA, they take into account a variety of training and testing ratios.

Table 3: All Model Performance for the 80% and 20% of training and testing ratio.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Without Using PCA					
SVM	79.02	74.43	78.03	71.33	87.22
NB	78.18	73.43	78.92	76.53	87.87
RF	83.65	86.98	75.65	80.02	77.94
DT	72.55	71.12	73.04	72.01	80.82
Without Using PCA					
SVM	86.08	88.88	86.90	88.65	92.91

NB	89.54	88.23	85.93	85.83	92.33
RF	89.86	89.18	89.77	89.91	93.72
DT	82.02	84.65	73.91	73.92	87.55

The current analysis evaluated the accuracy of the machine learning technique with the accuracy of methods utilized in three prior investigations, each using a different methodology. The results are shown in Table 4. The reference articles [40, 41] and [42] used the Levenberg-Marquardt methodology, the Genetic Algorithm with Radial Basis Function Neural Network (GA_RBF NN), and the Modified Particle Swarm Optimization Neural Network (MPSO-NN) approaches, respectively.

Table 4: Comparison between this study and previous studies

Reference	Method	Accuracy
[33]	Levenberg-Marquardt	82%
[34]	GA_RBF NN	77.4%
[35]	MPSO-NN	81.8%
This study	RF	89.86%

The research used the Random Forest (RF) classification algorithm, which differs from the approach utilized in the prior study. The accuracy results indicate that the RF strategy used in the current study achieved a much higher accuracy rate of 89.86% compared to the methodology used in previous investigations. The reported accuracy in this study surpassed that of the Levenberg-Marquardt method 82% accuracy in [40], the GA_RBF NN approach 77.4% accuracy in [41], and the MPSO-NN methodology 81.8% accuracy in [42].

5. Conclusion

The primary objective of our work was to construct appropriate categorization models to assist in the timely identification of diabetes. Although the Pima Indian Diabetes dataset included several missing variables, we successfully addressed these problems. PCA was used to select features and eliminate outliers in our data processing procedures, resulting in improved models.

By combining these techniques with classifiers such as Support Vector Machine, Random Forest, Naïve Bayes, and Decision Tree, we attained a remarkable accuracy rate of 89.86%. The training phase was deemed effective as it used 80% of the dataset. This unequivocally demonstrates that our methodology reliably identifies instances of diabetes. Enhancing the diagnosis of diabetes is essential for improving disease management and patient outcomes. Our research demonstrates that the use of sophisticated machine learning techniques is necessary to achieve this objective.

References

[1] A. Salih, S. T. Zeebaree, S. Ameen, A. Alkhyyat, and H. M. Shukur, "A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection," in *2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)*, 2021: IEEE, pp. 61-66.

[2] D. A. Hasan, S. R. Zeebaree, M. A. Sadeeq, H. M. Shukur, R. R. Zebari, and A. H. Alkhayyat, "Machine learning-based diabetic retinopathy early detection and classification systems-a survey," in *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, 2021: IEEE, pp. 16-21.

[3] B. R. Ibrahim *et al.*, "Embedded system for eye blink detection using machine learning technique," in *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, 2021: IEEE, pp. 58-62.

[4] I. Zeebaree, "The Distributed Machine Learning in Cloud Computing and Web Technology: A Review of Scalability and Efficiency," *Journal of Information Technology and Informatics*, vol. 3, no. 1, 2024.

[5] H. I. Dino, S. R. Zeebaree, D. A. Hasan, M. B. Abdulrazzaq, L. M. Haji, and H. M. Shukur, "COVID-19 diagnosis systems based on deep convolutional neural networks techniques: A review," in *2020 International Conference on Advanced Science and Engineering (ICOASE)*, 2020: IEEE, pp. 184-189.

[6] B. K. Hussan, Z. N. Rashid, S. R. Zeebaree, and R. R. Zebari, "Optimal Deep Belief Network Enabled Vulnerability Detection on Smart Environment," *Journal of Smart Internet of Things*, vol. 2022, no. 1, pp. 146-162, 2023.

[7] L. M. Abdulrahman, A. M. Abdulazeez, and D. A. Hasan, "COVID-19 world vaccine adverse reactions based on machine learning clustering algorithm," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 134-140, 2021.

[8] B. W. Salim, B. K. Hussan, Z. S. Ageed, and S. R. Zeebaree, "Improved Transient Search Optimization with Machine Learning Based Behavior Recognition on Body Sensor Data," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 75, no. 2, pp. 4593-4609, 2023.

[9] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare," in *2018 fourth international conference on computing communication control and automation (ICCUBEA)*, 2018: IEEE, pp. 1-6.

[10] K. G. M. M. Alberti, P. Zimmet, and J. Shaw, "International Diabetes Federation: a consensus on Type 2 diabetes prevention," *Diabetic Medicine*, vol. 24, no. 5, pp. 451-463, 2007.

[11] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578-1585, 2018.

[12] W. H. Organization, *World Health Statistics 2016 [OP]: Monitoring Health for the Sustainable Development Goals (SDGs)*. World Health Organization, 2016.

[13] M. Franciosi *et al.*, "Use of the diabetes risk score for opportunistic screening of undiagnosed diabetes and impaired glucose tolerance: the IGLOO (Impaired Glucose Tolerance and Long-Term Outcomes Observational) study," *Diabetes care*, vol. 28, no. 5, pp. 1187-1194, 2005.

[14] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS international conference on computer systems and applications*, 2008: IEEE, pp. 108-115.

[15] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert systems with applications*, vol. 33, no. 4, pp. 847-856, 2007.

[16] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung journal of medical sciences*, vol. 29, no. 2, pp. 93-99, 2013.

[17] N. Abdulhadi and A. Al-Mousa, "Diabetes detection using machine learning classification methods," in *2021 international conference on information technology (ICIT)*, 2021: IEEE, pp. 350-354.

[18] N. K. Putri, Z. Rustam, and D. Sarwinda, "Learning vector quantization for diabetes data classification with chi-square feature selection," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 546, no. 5: IOP Publishing, p. 052059.

[19] T. Nadira and Z. Rustam, "Classification of cancer data using support vector machines with features selection method based on global artificial bee colony," in *AIP conference proceedings*, 2018, vol. 2023, no. 1: AIP Publishing.

[20] Z. Rustam, J. Pandelaki, and A. Siahaan, "Kernel spherical k-means and support vector machine for acute sinusitis classification," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 546, no. 5: IOP Publishing, p. 052011.

[21] T. V. Rampisela and Z. Rustam, "Classification of schizophrenia data using support vector machine (SVM)," in *Journal of Physics: Conference Series*, 2018, vol. 1108: IOP Publishing, p. 012044.

[22] M. S. Islam, M. K. Qaraqe, and S. B. Belhaouari, "Early prediction of Hemoglobin Alc: a novel framework for better diabetes management," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020: IEEE, pp. 542-547.

[23] G. Pethunachiyar, "Classification of diabetes patients using kernel based support vector machines," in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020: IEEE, pp. 1-4.
[24] O. M. Alade, O. Y. Sowunmi, S. Misra, R. Maskeliūnas, and R. Damaševičius, "A neural network based expert system for the diagnosis of diabetes mellitus," in *Information technology science*, 2018: Springer, pp. 14-22.
[25] S. Ravizza *et al.*, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nature medicine*, vol. 25, no. 1, pp. 57-59, 2019.
[26] R. R. Asaad and R. M. Abdulhakim, "The Concept of Data Mining and Knowledge Extraction Techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 17-20, 2021.
[27] M. Padmavathi and C. Sumathi, "A New method of data preparation for classifying diabetes dataset," *Indian Journal of Science and Technology*, vol. 12, no. 22, pp. 1-9, 2019.
[28] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
[29] A. H. B. Aighuraibawi *et al.*, "Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm," *Comput. Syst. Sci. Eng.*, vol. 47, no. 1, pp. 553-574, 2023.
[30] A. A. Farabe, T. S. Sharika, N. Raonak, and G. Ashraf, "A supervised learning approach by machine learning and deep learning algorithms to predict type II DM risk," Brac University, 2019.
[31] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.
[32] B. Selvaraj, S. Pavithra, A. N. Rak, and M. Jeyaselvi, "Prediction and Detection of Diabetes Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, pp. 2395-0056, 2020.
[33] D. A. Zebari, S. S. Sadiq, and D. M. Sulaiman, "Knee osteoarthritis detection using deep feature based on convolutional neural network," in *2022 International Conference on Computer Science and Software Engineering (CSASE)*, 2022: IEEE, pp. 259-264.
[34] S. Rukhsar *et al.*, "Artificial intelligence based sentence level sentiment analysis of COVID-19," *Computer Systems Science and Engineering*, vol. 47, no. 1, pp. 791-807, 2023.
[35] K. L. Priya, M. S. C. R. Kypa, M. M. S. Reddy, and G. R. M. Reddy, "A novel approach to predict diabetes by using Naive Bayes classifier," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 2020: IEEE, pp. 603-607.
[36] D. A. Zebari, D. M. Sulaiman, S. S. Sadiq, N. A. Zebari, and M. S. Salih, "Automated Detection of Covid-19 from X-ray Using SVM," in *2022 4th International Conference on Advanced Science and Engineering (ICOASE)*, 2022: IEEE, pp. 130-135.
[37] H. Zhang, "The optimality of naive Bayes," *Aa*, vol. 1, no. 2, p. 3, 2004.
[38] K. I. Taher, A. M. Abdulazeez, and D. A. Zebari, "Data mining classification algorithms for analyzing soil data," *Asian Journal of Research in Computer Science*, vol. 8, no. 2, pp. 17-28, 2021.
[39] S. CHALO and İ. B. AYDİLEK, "A New Preprocessing Method for Diabetes and Biomedical Data Classification," *Qubahan Academic Journal*, vol. 2, no. 4, pp. 6-18, 2022.
[40] N. I. Alghurair, "A Survey Study Support Vector Machines and K-MEAN Algorithms for Diabetes Dataset," *Academic Journal of Research and Scientific Publishing| Vol*, vol. 2, no. 14, 2020.
[41] D. K. Choubey and S. Paul, "GA_RBF NN: a classification system for diabetes," *International Journal of Biomedical Engineering and Technology*, vol. 23, no. 1, pp. 71-93, 2017.
[42] K. Ateeq and G. Ganapathy, "The novel hybrid Modified Particle Swarm Optimization–Neural Network (MPSO-NN) Algorithm for classifying the Diabetes," *International Journal of Computational Intelligence Research*, vol. 13, no. 4, pp. 595-614, 2017.