Qubahan Academic Journal
گوڤارا قبهان یا ئەکادیمی
مجلة قبهان الأكاديمية
Doi: 10.48161/Issn.2709-8206

A New Preprocessing Method for Diabetes and Biomedical Data Classification

Sarbast CHALO[1,*]
Harran University, Engineering Faculty, Department of Computer Engineering, Şanlıurfa, Turkey

İbrahim Berkan AYDİLEK[2]
Harran University, Engineering Faculty, Department of Computer Engineering, Şanlıurfa, Turkey

* Corresponding author: Sarbast CHALO[1], sarbastchalo559@gmail.com

***Abstract*- People of all ages and socioeconomic levels, all over the world, are being diagnosed with type 2 diabetes at rates that are higher than they have ever been. It is possible for it to be the root cause of a wide variety of diseases, the most notable of which include blindness, renal illness, kidney disease, and heart disease. Therefore, it is of the utmost importance that a system is devised that, based on medical information, is capable of reliably detecting patients who have diabetes. We present a method for the identification of diabetes that involves the training of the features of a deep neural network between five and 10 times using the cross-validation training mode. The Pima Indian Diabetes (PID) data set was retrieved from the database that is part of the machine learning repository at UCI. In addition, the results of ten-fold cross-validation show an accuracy of 97.8%, a recall OF 97.8%, and a precision of 97.8% for PIMA dataset using RF algorithm. This research examined a variety of other biomedical datasets to demonstrate that machine learning may be used to develop an efficient system that can accurately predict diabetes. Several different types of machine learning classifiers, such as KNN, J48, RF, and DT, were utilized in the experimental findings of biological datasets. The findings that were obtained demonstrated that our trainable model is capable of correctly classifying biomedical data. This was demonstrated by achieving higher 99% accuracy, recall, and precision for parikson dataset.**

Keywords- Biomedical dataset, Diabetes, Machine learning, Classification.

I. INTRODUCTION

The protection and prevention of people in the community from diseases that are a risk to their health is the primary focus of public health efforts. Programs like immunization have assisted individuals in living longer lives, which is one of the primary reasons why governments invest a considerable amount of their GDP in public health and safety [1]. Despite this, chronic and inherited diseases that have a negative impact on public health have been on the rise for a number of years now. Diabetes mellitus is among the most fatal diseases because it can cause a wide range of complications, including injury to the kidneys, the heart, and the nerves [2]. Chronic hyperglycemia is one of the symptoms of diabetes. Other symptoms include problems in the metabolism of carbohydrates, lipids, and proteins owing to insulin deficiency or insulin resistance. Both of these conditions are caused by insulin resistance. Diabetes is a condition that lasts a long time because it causes high quantities of sugar to be found in the blood [3]. The majority of people in the globe have type 2 diabetes, which affects around 90–95 percent of the population. The prevalence of diabetes is increasing at a startlingly rapid rate. Complications related to diabetes are a leading cause of death, and the disease itself is responsible for the loss of thousands of lives every year without anybody noticing. By the year 2035, 592 million people will have diabetes, which is more than twice as many as are currently impacted by the disease [4]. When the glucose level in the blood is too high, it causes damage to the kidneys, heart, brain, and small blood vessels in the eyes. When it comes to the practice of medicine, making a diagnosis is among the most challenging and significant activities. It is possible to forecast whether or not a person will develop diabetes based on data obtained from patients. A few examples of these include the amount of glucose in the plasma, the diastolic pressure, the thickness of the bicep's skinfold, the serum insulin level, the body mass, and the age. After that, an appointment with a specialist is made for the patient. When making a decision, the attending physician will take each of these criteria into account [5].

Early diagnosis involves a physician using his or her expertise and experience to make a prognosis and diagnose a disease. Nevertheless, this method is not foolproof and may produce false results. The healthcare industry accumulates a vast amount of data linked to healthcare, but that data is unable to recognize previously unseen trends, which makes it difficult for the industry to make informed judgments. Because manual judgments are based on the observations and judgments of a healthcare professional, which is not always reliable [6], they can pose a significant risk when it comes to the early detection of diseases. There may be certain hidden patterns that continue to exist, which may have an effect on the observations and outcomes. As a consequence of this, patients are receiving services of poor quality. Hence, a sophisticated mechanism is necessary for the early detection of disease, preferably one that features an automated diagnosis and improved accuracy. A wide variety of faults that go unnoticed and patterns that are kept hidden give rise to a diversified set of machine algorithms and data mining techniques that can produce accurate and dependable results. The day-to-day effects of diabetes are becoming increasingly significant, which has led to the development of a number of data mining algorithms that can extract hidden patterns from enormous amounts of healthcare data. In addition, such data can be put to use for the purposes of feature selection and the automated prediction of diabetes [7].

The decision-making process could take weeks or months, which makes the job of the physician difficult because of the length of time involved. Because large amounts of medical data are now easily accessible, it is now feasible to perform research in a wide number of subfields that fall under the umbrella of the medical sector. As a consequence of this, it is extremely challenging, if not downright impossible, for a human to manage such enormous amounts of data. Therefore, computer-based strategies have supplanted traditional procedures in favor of ones that are more efficient. The use of computerized procedures not only increases precision and productivity but also lowers associated expenses. Deep learning is a rapidly developing idea that operates in a manner that is strikingly similar to that of the human brain. Multiple levels of data representation and selection invariance resolution that is both efficient and effective [8]. Deep learning algorithms are utilized in a wide difference of applications within the field of medical diagnosis and prognosis. According to a substantial body of literature, deep learning algorithms generate greater outcomes, have lower classification error rates, and are more immune to noise. It is able to process vast volumes of data and decipher even the most complicated of situations with a relatively simple level of effort [9]. Some recent medical diagnoses have made use of techniques such as machine learning, bioinspired computing techniques, and deep learning methodologies [10]. To ameliorate the accuracy of our classification of diabetes mellitus, we made use of deep neural networks, a method that has seen a surge in popularity in recent years within the field of machine learning. Furthermore, the dataset is trained properly before it has been utilized to diagnose diabetes; consequently, the test dataset delivers a correct result in the vast majority of situations. So far, it has not been possible to successfully predict diabetes using methods based on machine learning [11]. On the other hand, the accuracy rate of our method is far higher than that of the current state of the art. There have been a lot of different approaches taken, but none of them have been able to deliver results that are reliable and consistent.

II. LITERATURE REVIEW

The procedures that were previously used have been superseded by the techniques of data mining, which feature superior classification, accuracy, and precision. In addition, machine learning is a technique of artificial intelligence that discovers links between nodes without the need for the nodes to be trained beforehand [12]. The primary advantage of using machine learning approaches is that they may drive prediction models without requiring intensive prior training in relation to the underlying mechanisms. Data mining and machine learning are two methods that can be used to uncover data that has been hidden by an innovative strategy [13]. In this part of the article, we will examine some prior research to demonstrate the viability of the notion of data mining approaches to be used in the driving prediction model, mostly for diabetics.

The AdaBoost method was proposed as a potential solution by Sajida Perveen et.al, [14]. An AdaBoost ensemble model performs noticeably better than bagging and J48 when it comes to the classification of patients who have been diagnosed with diabetes. The author has been inspired to write this book by the ever-increasing impact that diabetes is having in each and every part of the planet. As a direct consequence of this, the diagnosis and avoidance of diabetes mellitus are gaining more and more importance in the world of medicine [15]. The author provided a classification algorithm that has demonstrated efficiency for the categorization of people who have diabetes in the community of Canada throughout all three age groups. On the test data, three distinct ensemble models, including bagging, AdaBoost, and J48, were utilized in order to evaluate how precise and effective the performance was. According to the results of the study, AdaBoost performs far better than its rivals do in terms of accuracy. According to the authors of the study, AdaBoost possesses the capability of enhancing disease prediction for a wide range of ailments, some of which include coronary heart disease and hypertension, amongst others. The primary roles that machine learning plays in the medical diagnosis and treatment of various diseases Alade et al. [16] introduced a framework for the classification by developing artificial neural networks (ANN) and Bayesian networks, including a multiple ANN architecture that provides a back-propagation technique as well as a Bayesian regulation method for training and testing the dataset. This method was able to successfully predict diabetes. For the purpose of evaluating the dataset, this technique was presented as both a back-propagation technique and a Bayesian regulation algorithm. Using this procedure, it was possible to determine whether or not a patient will go on to develop diabetes. The data have been trained in such a way that the regression graph will

display the results in an appropriate manner. With the help of this model, a diagnosis may be carried out from a distance, and the system is able to communicate with patients even when it is not there in the room with them.

Diabetes was classified by N. K. Putri, Z. Rustam, and D. Sarwinda using Learning Vector Quantization and Chi-Square feature selection [17]. They achieved a classification accuracy of 100% despite only utilizing 80% and 90% of the available training data. K-Nearest Neighbor (KNN) is the other method that has also been used with this dataset, and it produced an accuracy of 91% [18]. Getting the appropriate k parameters might be difficult when using the KNN method, which is one of the approach's drawbacks. High k values mitigate the impact of background noise on classification but at the expense of less defined class boundaries. Conversely, k values that are too low result in fewer comparison samples, which in turn results in a reduction in accuracy [19]. T Nadira and Z Rustam classified cancer using SVM and feature selection, and the accuracy of their results was 99.9999% on lung cancer datasets and 96.4286% on breast cancer datasets, respectively [20]. For the purpose of acute sinusitis categorization, SVM was utilized by Ariana, Z. Rustam, J. Pandelaki, and A. Siahaan, and they achieved an accuracy of 90% [21]. Z. Rustam and Rampisela T. V. also utilized SVM for the purpose of classifying data related to schizophrenia, and they achieved an accuracy of 90.1% [22].

III. PROPOSED METHOD

In order to accomplish what has to be done, the diabetes illness prediction system that has been presented involves linking together a number of different steps. The first step of our proposed method is data collection, this work collected different datasets namely, PIMA diabetic dataset and biomedical dataset including heart-c, heart-h, Indian liver, vertebral-column-2c, vertebral-column-3c, Parkinson's, and thyroid. The second step of our method is used to preprocess our dataset through different stages including data cleaning, filling in missing values, data resampling, and normalization. After this step, we obtained preprocessed dataset. In the third step, we used Principal Component Analysis (PCA) algorithm to obtain the most significant features. Then, data partitioning is performed based on using the K-fold cross-validation method. After that, we utilized the ML approach to train the system with the most optimal parameters based on the training data. The models that have been trained will be able to make accurate predictions based on the test samples. We have used different classifiers which are KNN, J48, RF, and DT. Finally, we evaluate our proposed method using different evaluation matrices which are accuracy, precision, recall, and F1-measure. The obtained results were then compared with several previous studies in the literature. The comprehensive system flowchart that has been presented in this study is given in Figure 1.

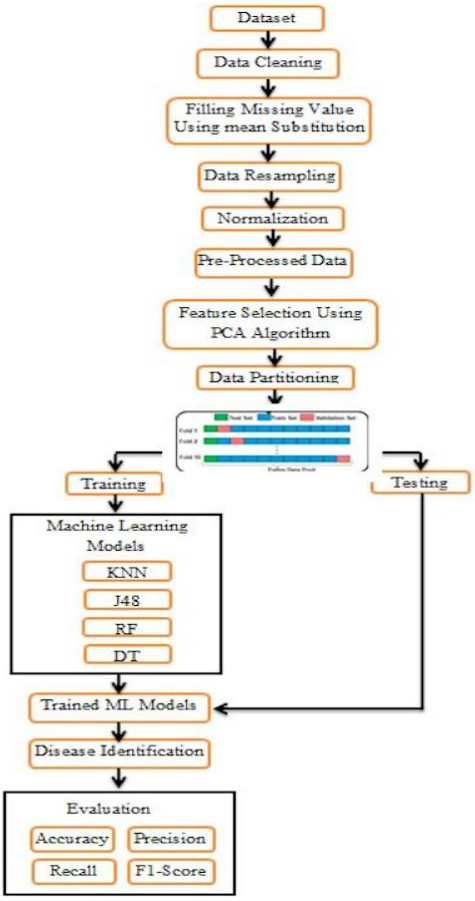


Figure 1: Proposed System Flowchart

a) Dataset

1) PIMA

A dataset that was received from Kaggle was utilized for the purposes of this investigation in this work, an online database of datasets (https://www.kaggle.com/uciml/pima-indians-diabetes-database), and it was utilized in the research process. This dataset has a number of attributes and based on them, we were able to make a prediction as to whether or not the patient would get diabetes. The dataset only contains cases of women who are at least 21 years old, and all of them are female. There is a total of 768 people included in the dataset; of them, 268 samples have been determined to have diabetes, while the remaining 500 samples have been determined to not have diabetes. The dataset is comprised of nine different characteristics, which are as described in the following: the number of pregnant women, the serum glucose presence, the diastolic, the insulin levels, the body mass index (BMI), the triceps, the skinfold thickness, the Mellitus purebred feature, age, and a category variable. Additionally, the dataset contains a class variable. On the other hand, the other eight characteristics are what are known as "feature variables" or "independent variables. There are only two possible values for the class variable: yes and no. A value of "Yes" indicates that the person has diabetes, whereas a value of "No" indicates that they do not have diabetes. The parameters of the dataset are broken down into further specifics in Table 1, which can be found here [20].

Table 1. Description of the dataset attributes.

Sr. No.	Parameters of Dataset	Description of Parameters	Normal
1	Number of Pregnancies	Number of times the person gets pregnant	0 - 17
2	Plasma glucose concentration	Repeats the concentration of glucose in a person's body.	0 - 199
3	Diastolic blood pressure	Represent the diastolic blood pressure in (mm Hg)	0 - 122
4	Triceps skinfold thickness	Represent triceps skinfold thickness in (mm)	0 - 99
5	Serum insulin	Represent 2 h serum insulin in (μU/mL)	0 - 846
6	Body mass index	It is a value derived from the weight and height of a person (weight in kg/(height in m)2)	0 – 67.1
7	Diabetes pedigree function	It represents the history of diabetes associated with a particular person.	0.0078 – 2.42
8	Age	Age of the person in years	21 - 81
9	Class variable	It represented two classes: diabetic and non-diabetic	Yes/No

2) Biomedical Dataset

The machine learning repository at UCI provides access to a wide range of datasets that may be utilized for research. This repository also allows users to access the many different types of biological datasets that are kept inside it. The biomedical datasets are retrieved from the repository at UCI over the course of the inquiry, and they are shown in Table 2. Both the Cleveland hospital and the Hungary Institute of Cardiology contributed data to the heart disease datasets that are known as heart-c and heart-h, respectively. These datasets were named after their respective initials. Examples of features include a reading of the patient's blood pressure when they are at rest, the amount of serum cholesterol, the patient's age, gender, and type of chest pain [23]. The topic of the class attribute is either the existence of heart disease in the patient or the absence of heart disease in the patient, depending on which one of those two possibilities is true. One sample of liver tissue taken from an Indian individual has the ability to disclose 10 distinct aspects. This formula takes into account a variety of factors, some of which include total proteins, albumin, the age and gender of the patient, total bilirubin, direct bilirubin, and Alkphos. Other factors include SGPT and SGOT. The records that are stored in databases make it possible to determine whether or not a patient suffers from liver disease.

The orthopedics field is intended to make use of the dataset that contains information on individuals who have disorders affecting their spinal column. Other datasets classify individuals into normal or pathological groups based on their lumbar lordosis angle, sacral slope, and degree of spondylolisthesis. Additionally, the pelvic incidence, tilt, and radius of an individual's pelvis are also taken into consideration. Patients are classified as normal, having a disk hernia, or having spondylolisthesis when using this technique to categorize. There are three categories in total. The Parkinson's dataset is made up of biomedical voice measures that were taken from a total of 31 people, 23 of whom have been diagnosed with Parkinson's disease. There are around six voice records connected with the medical file of each individual patient. Every recording will have an average, maximum, and minimum voice fundamental frequency, as well as various measurements of variation in fundamental frequency and amplitude, among other things. This information will be included in the file. The recording can be analyzed in a number of different ways depending on how this information is used. The information gathered from a patient's thyroid gland is utilized to assess whether or not the patient's thyroid gland is operating normally, hypofunctioning, or hyperfunctioning. [24] Triiodothyronine, a basal level of thyroid-stimulating hormone (TSH), changes in TSH rate following injection of 200 micro grams of thyrotropin-releasing hormone, and T3-resin uptake test are the characteristics that are included.

Table 2: Biomedical Datasets

Dataset name	Number of Attributes			
	Nominal	Continuous	Class	Record
Heart-c	6	7	5	303
Heart-h	6	7	5	294
Indian-liver	1	9	2	579
Vertebral-column-2c	0	6	2	310
Vertebral-column-3c	0	6	3	310
Parkinsons	0	22	2	195
Thyroid	0	5	3	215

b) Data Preprocessing

The first step that must be done in order to analyze the dataset is to clean it. This involves using a methodical technique to remove of records and attributes that are not needed. To start, the dataset contains a number of categorical values that must be removed so as not to compromise patients' privacy. These include the hospital number, the event date, and the episode description. In addition, the dataset lacks values for some patients' diabetes types, which is a crucial piece of information for our study due to the fact that we investigated diabetes complications in diabetic patients. This information is essential because of the nature of the study.

In the second step of preprocessing, when training classifiers, dealing with missing values is critical because the majority of the existing ML techniques are not able to be used with data that is missing values. Only the nationality variable in our dataset contains categorical values, and it is this attribute that is causing these kinds of problems. Therefore, the values that were absent were filled in with the value that appeared in that column the most frequently. Mean replacement is a statistical method that simply entails representing and filling any missing value in a characteristic with the average of observational data for that characteristic in other files. This is done in order to guarantee that the data is comprehensive.

In the third step of preprocessing, dealing with unbalanced datasets is one of the most typical issues that arise during the process of constructing and training data mining models using ML. This issue must be overcome when developing and training ML models. Because our dataset contains examples of this issue, resampling has been implemented as a solution to address the issue. It is clear that the majority of the problems are out of whack with one another. To be more specific, the neuropathy, nephropathy, retinopathy, and diabetic foot characteristics are all characterized by significant imbalances in their distributions. For example, diabetic foot only occurs in 2.5% of the entire number of records that have been compiled. It is necessary to find a solution to this problem that is not only powerful but also economical in order to achieve a satisfactory level of harmony. One option would be to use the under-sampling method in order to lower the number of samples in the class that makes up the majority.

In the final step of preprocessing, we were able to do feature scaling, which enhanced the time processing of ML algorithms with the assistance of normalizing the data within the range [0–1]. By increasing the amount of time spent digesting our proposed model, we are able to accomplish the goals of this process. The process of rescaling the characteristics in order to achieve a typical normal distribution with zero mean and unit variance is referred to as Z-score normalization. As demonstrated in (4), the process of standardization (R) also has the effect of reducing the skewness of the data distribution.

$$R(x) = \frac{x - \bar{x}}{\sigma} \qquad (1)$$

when x is the occurrence of the feature set that has n-dimensional, and x ∈ Rn.$\bar{x}$ ∈ Rn and σ ∈ Rn are the standard deviation and mean of the characteristics. In contrast, several ML methods, including tree-based models, are undoubtedly instances of models for which feature standardization could ensure a meaningful increase. This is the case for many ML models. Following the completion of the preprocessing stage, we have a total of 698 instances, 467 of which do not have diabetes, while 232 of the cases do have diabetes. As a result, the data set is improved by this section, and the subsequent sections will be better equipped to take advantage of the improvements.

Before the data on Pima Indians and diabetes types can be accepted by a machine learning network, the data on those populations must first be preprocessed. If training and testing datasets are kept separate, models will only learn from the training data and will have their performance evaluated using the testing data. The data was split into two sets: the training set and the testing set.

c) Principal Component Analysis (PCA)

PCA is largely based on a statistical–mathematical technique that makes use of an eigenvector for the aim of decreasing the size of a record set [45]. This approach was developed in order to reduce the number of records that need to be analyzed. This strategy takes into account multiple variables, each of which is connected to a distinct retaining and the highest variance within the dataset. PCA is a statistical technique that is based on the orthogonal components of linear datasets. It is also known as the orthogonal component analysis (OCA). The first component of the PCA method is comprised of a group of variables and linear data that, for the most part, is made up of information that is connected to the variables. The score, which is determined using a scale ranging from 0 to 100 and is used to quantify quality, is inversely related to the quality. As a result, a better quality can be inferred from a score that is higher. In principal component analysis (PCA), the dataset that is used has to be scaled, and the method itself provides a summary of the impacts of producing statistics, which might be especially sensitive to proportional measurements

d) Training Method

The subsequent phase, which followed the processing of the dataset and the selection of the ML techniques that would be used, was the phase of actually developing the frameworks by training each method utilizing the dataset that has been preprocessed. This phase followed the previous phase of choosing the ML algorithms that would be utilized. After the machine learning algorithms that were going to be utilized had been chosen, this step was taken. Extensive testing was done in order to both "train" the models and "fine-tune" them in preparation for the final presentation. In the paragraphs that follow, we shall examine each stage in greater depth as we move forward.

1) Cross-validation

The k-fold cross-validation (KCV) procedure is one of the methods that is utilized most commonly to select a classifier and assess the effectiveness of that predictor. This method is one of the ways that was developed in the 1970s. Figure 2 is a pictorial depiction of the data splitting methodology, which includes all of the visual material that is necessary to understand the method (with tenfold cross-validation). In order to do the analysis, the dataset was divided into K folds. K-1 folds were deployed in to train and fine-tune the hyperparameters in the inner loop, which is where the grid search method [25] was performed. This was accomplished by the utilization of the K-1 folds. The model was evaluated using the test data as well as the hyperparameters that were determined to be the most accurate. Because the records in the dataset are not distributed evenly, stratified KCV [26] was applied in order to maintain the same proportion of samples in each class that was present in the dataset's initial percentage. In addition, in order to get a more accurate assessment, we repeated this procedure another ten times. It was determined how to calculate the final performance metric by utilizing Equation 6where M is the overall performance measure for the predictor and Pn denotes the performance metric for each fold, where n can range from 1 to K, the ultimate performance measure being M.

$$M = \frac{1}{K} \times \sum_{n=1}^{K} P_n \qquad (2)$$

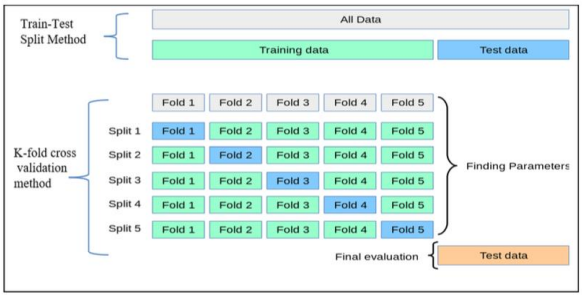


Figure 2: K-fold cross-validation method and train-test splitting method

2) Classification

The classification model is dependent on the aforementioned components to obtain the training data and the testing data. The following are the machine learning techniques that will be used to handle the training data: K-Nearest Neighbor, J48, Random Forest, and Decision Tree (DT). The results indicate that SVM is capable of achieving the best possible outcome for categorization data. In contrast, in order to perform classification based on the best parameters that originate from the machine learning algorithms, the testing data also needs to pass in the current part of the process. The process of the suggested model is depicted here in Figure 3, which can be found below.

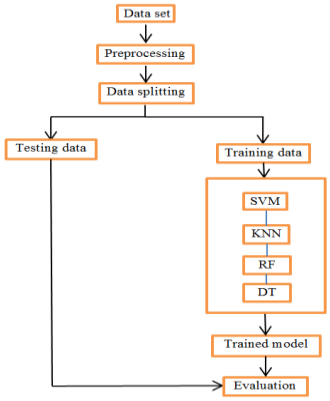


Figure 3: Proposed Classification Framework

J48 classifier

The C4.5 method is a technique that can be used in classification that generates decision trees by adhering to the guidelines established by information theory. It is an extension of Ross Quinlan's older ID3 technique, which is also known as J48 in Weka, where J stands for Java. This new approach was developed by Ross Quinlan. C4.5 is usually referred to as a statistical classifier because of its capacity to generate decision trees that may be used for categorization. One of the causes behind this is as follows: The J48 version of the C4.5 algorithm offers a great deal of supplementary functionality. This capability includes the ability to derive rules, as well as take into consideration continuous attribute value ranges, missing value support, and the pruning of decision trees, among other things. The WEKA data mining tool has what is known as an open-source Java implementation of the C4.5 method. This implementation is given the moniker J48.This implementation is written in Java. Users of J48 have the option of classifying data utilizing either decision trees or rules that are derived from those trees [48].

Utilizing the idea of information entropy, this method, which is quite analogous to the ID3 algorithm, constructs decision trees on the basis of a collection of training data. This procedure is performed three times in total. The instance utilized for training come in the form of a set denoted by the notation S='s1, s2,...'. These samples have already been categorized. Each sample si is characterized by a p-dimensional vector that includes the components x1, x2, xp, and I in its make-up. These elements are a reflection of the attribute values or characteristics of the associated sample, as well as the category in which the sample is categorized. Additionally, these elements represent the sample's placement. If you want the classification accuracy to be as good as it possibly can be, then you should make the divides based on the property that includes the most information. The C4.5 method analyzes the data at each node of the tree to decide which data characteristic would most efficiently split its group of samples into subgroups that are enriched in one class or the other. These subsets are able to have their respective classes enhanced. The criterion for splitting is the normalized information gain, and it is determined by computing the difference in entropy that exists between the two groups. The choice is determined on the basis of the property that presents the biggest possible opportunity for an increase in the amount of normalized information. Using a divide-and-conquer strategy, the C4.5 algorithm will initially perform recursive operations on the partitioned sub lists before moving on to the construction of a decision tree based on the greedy algorithm. The algorithm is capable of producing the decision tree as a result of this. This method is repeated a great number of times till the desired result has been achieved. In an even more concrete and high-level example, the algorithm performs on the decision tree seen in Figure 1 by looking at the data that have already chosen to take said classes, and it utilizes that data to construct a model to detect diabetic. An algorithm that works on the decision tree shown in Figure 1. To begin, we will take our list of students and divide them into a variety of data sets using a random selection process. After that, we produce, for each data set, a set of weights that makes a prediction about the path that a student will take. Throughout the end, we choose the data collection that can produce the most accurate prediction regarding the path that a student will pursue in their academic career.

Decision tree

A decision tree method, also known as a DT classifier, is a decision-making aid that takes the form of a tree structure and is built with the help of features that are entered [26]. The primary goal of this particular classifier is to construct a model that makes predictions about the target variables based on a number of input features. Because it is simple to derive decision rules from any given set of input data, this classifier is applicable to a wide variety of different kinds of applications [27, 29].

The DT approach is a method for nonparametric supervised learning that may be utilized for solving issues involving regression and classification. Figure 4 demonstrates how the DT model might be seen as a representation of a structure. The root node, the division, and the leaf node are the three nodes that make up this paradigm. Every single internal node is actually just a test that's being run on some property. Each division is the outcome of that test, and each leaf node stores the class label for its parent. The root node is the point at which the construction of the tree actually begins. At first, an attribute is selected to be put at the root node, where it will stay for the duration of the process. After that, a division is performed for each of the potential values. This results in the creation of subgroups inside the dataset, one for each of the possible values that may be found in the property. The process of the tree is carried out in a recursive way for each division, and the cases that reach the branch are the only ones that are taken into consideration. It is possible to halt the progression of the tree when all of the cases on a node have the same categorization. When attempting to determine the optimal tree partition, entropy or classification error are typically the two metrics of choice [30].

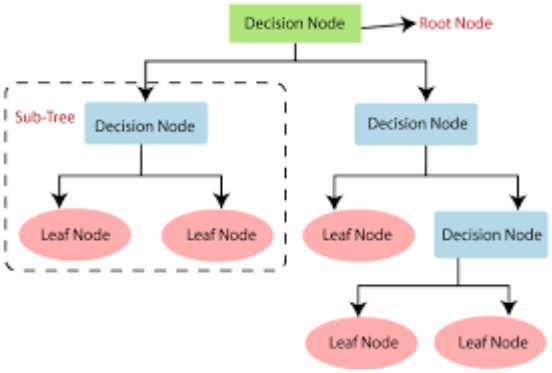


Figure 4: Decision Tree Algorithm [30]

K-Nearest Neighbor (KNN)

A classifier known as K-Nearest Neighbor, which also goes by the name instance-based classifier, is a member of the family of learning strategies known as lazy learning. Classification is able to be accomplished with the help of the K-NN classifier. However, it excels at solving problems involving categorization the most. The distance between the test data point, also named the query data point, and each of the other data points in the training set is evaluated by the classifier when using the KNN algorithm. After this, it locates the K data points that are the closest neighbors to the current test data point. The predicted category is then selected through a straightforward vote process using the K data points that are geographically closest to one another. It is a non-parametric, slow-learning algorithm that does not learn anything during the training phase. The method is straightforward when applied to a large dataset. In most cases, a bigger k number indicates that there will be less of an effect that noise has on the findings. If you use all of a dataset's training samples to generate each query instance, you will notice a considerable increase in the amount of time it takes to perform the calculations [31].

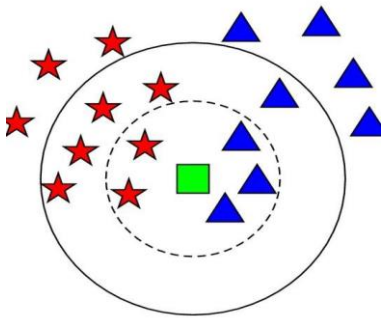


Figure 5: K-Nearest Neighbor Algorithm [31]

Random Forest

RF is all about an ensemble of models; each model forecasts an outcome, and in the end, the model with the majority of correct predictions wins. RF is a group of models that can be played together like an orchestra. It has been examined in the research that has been done on the subject, and it has been demonstrated to be effective in predicting diabetes. A Random Forest is a flexible learning model that can solve both regression and classification problems. During the training phase, Random Forest constructs a large number of "decision trees," and then it generates a forecast that is an average of all of the forecasts generated by the decision trees. Both categories of challenges are within the scope of this model's capabilities [32]. In regression, the aim variable is continuous, but in classification problems it is categorical. Regression uses continuous variables. In order to achieve a high level of accuracy using Random Forest, EDA approaches are utilized. Combining a large number of relatively weak models results in the production of a more robust model. The algorithm known as "Random Forest" has the capacity to process large datasets that have a high level of spatiality. The ability of the model to process a vast number of input variables and determine which ones are the most significant is referred to as the model's dimensionality reduction capability. A useful characteristic of the model is that it highlights the relevance of variables even when working with a random dataset. In contrast to the conventional model, which uses a voting strategy that is comparable to one another, the "Adaptive Random Forest" (ARF) is the model that performs the best when an unequal voting strategy is used [33].

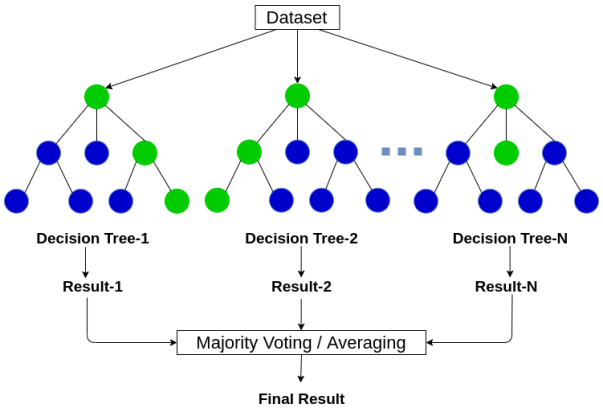


Figure 6: Random Forest [33]

IV. Experimental Results

Several various test metrics were utilized to put the constructed models through their paces and determine how well they performed. The classification accuracy measure is the first one that is utilized. Classification accuracy is known as the percentage of occurrences that are correctly labeled according to their real class labels. Even though it is one of the most common assessment measures, when applied to datasets that are not evenly distributed, it does not provide a true description of the performance of the model. As a result, it is essential to make use of many alternative strategies in this scenario. It is possible to calculate the accuracy of a classifier by using Equation 3, in which TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative. The term "sensitivity" refers to the proportion of instances that are correctly classified as positive as well as all positive instances given in Equation (4). Specificity is calculated based on the percentage of correctly labeled negative cases as well as the total number of negative instances stated in Equation (5).

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

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

$$Specificty = \frac{TN}{FP+TN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

a) Experimental Results Using PIMA and Biomedical Dataset

In this part, a comparative testing evaluation of all the traditional ML techniques has been performed for the purpose of classifying diabetes mellitus into positive and negative groups. It has been performed to compare and analyze the degrees of accuracy offered by the various traditional algorithms. Experiments have been conducted using two different datasets. One of these is the PIMA diabetic mellitus dataset, which is split into two classes: positive or negative. The second dataset is also a biomedical one, and it uses a binary classification system to divide the data into healthy and unhealthy categories.

In this study, a number of different machine ML such as KNN, J48, DT, and RF, were utilized to determine whether or not a person has diabetes based on a variety of datasets, such as pregnancies, glucose levels, blood pressure, and the thickness of their skin. Insulin levels were also taken into consideration. The dataset is divided into two sections: the training section and the testing section. The dataset is broken down as follows: 80% is comprised of training, while 20% is comprised of testing. The evaluation of the recommended system was conducted based on performance criteria such as accuracy, specificity, and sensitivity. Each makes use of their own unique machine learning algorithms, and the percentage results of each metric acquired by each methodology are as follows: The outcomes of this study are based on a number of different criteria for evaluation. A comparison with the machine learning techniques described in the previous section has also shown that our proposed model is superior. For the sake of this investigation, the dataset was divided into two subsets: one was used for training, and the other was utilized for testing. The dataset used for training had 547 diabetics and the dataset used for testing had 547 diabetics and 1053 non-diabetics, or 23% of all the data.

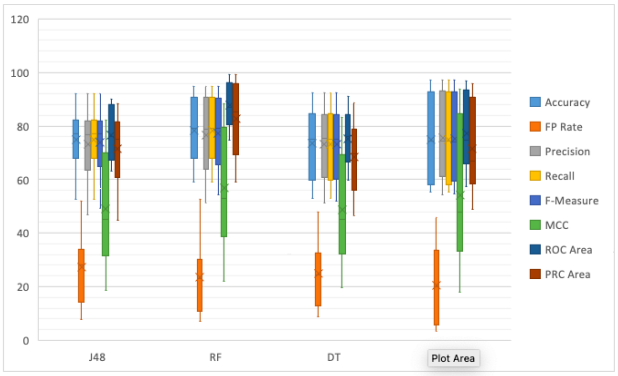
In this piece, we proposed a reliable method for predicting diabetes based on a machine-learning algorithm, which can determine whether or not a person has diabetes. We presented a comprehensive analysis that compared a number of different machine-learning approaches. The performance of these models was evaluated based on a number of different parameters, including accuracy, specificity, and sensitivity. The results of the evaluation indicate that the DNN strategy recommended is superior to all of the other approaches. In addition to this, we carried out an analysis that contrasted our method with the most recent developments in the field. To put it another way, a method of diabetes prediction using machine learning showed remarkable potential and performed far better when compared to the methods that are already in use. Utilizing this tactic can result in cost and time savings during the process of designing and developing diabetes illness prediction systems in the healthcare industry.

In experiment, a technique known as 10-fold cross-validation classification accuracy evaluation is used. The findings of the PIMA dataset after being put through 10 rounds of cross-validation using the machine learning classifier are presented in Table 3. Accuracy, sensitivity, and precision are some of the evaluation matrices that have been utilized in the process of performing performance evaluation. In different folds, the results that got the best accuracy, sensitivity, and precision were obtained by 70.18%, 70.2%, and 69.6%, respectively for KNN. When using J48, the results of different folds showed that 73.82% had the highest accuracy, while 73.8% and 73.5%, respectively, recall and precision. When using DT, the results of different folds showed that 70.96% had the highest accuracy, while 71% had obtained for sensitivity and precision. When using RF, the results of different folds showed that 75.26% had the highest accuracy, while 75.3% and 74.9%, respectively, had obtained for sensitivity and precision.

Then, other biomedical datasets are tested using same classifiers. The results are shown in Table 3. Table 3 presents the findings of each and every biological dataset evaluated with regard to its accuracy, recall, precision, and F-measure. In comparison, heart-c has the lowest accuracy success rate, which comes out to 52.8%, whereas thyroid has enough good results with the KNN classifier to obtain 97.2%. A higher recall of 97.2% was achieved for both the thyroid using KNN. The tests for Parkinson's disease, Vertabral-column-2c, and Vertabral-column-3c all had a specificity rate of one hundred percent. In the end, we were able to achieve a precision of 100% for both the Vertabral-column-2c and the Vertabral-column-3c.

Table 3: PIMA and Biomedical Dataset Classification Results (%) Using 10-Fold Cross-validation without Pre-processing

Dataset	Algorithm	Accuracy	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
diabetes	J48	73.8281	32.70	73.50	73.80	73.60	41.70	75.10	72.70
	RF	**75.2604**	31.40	**74.90**	**75.30**	**75.00**	**44.70**	**81.80**	**81.40**
	DT	70.9635	34.10	71.20	71.00	71.10	36.60	68.40	66.00
	KNN	70.1823	**37.80**	69.60	70.20	69.80	33.10	65.00	64.00
heart-c	J48	52.8053	**26.00**	47.00	52.80	49.40	28.10	63.10	44.80
	RF	**59.0759**	24.30	51.30	**59.10**	54.20	**37.10**	**80.00**	**59.10**
	DT	53.1353	20.20	51.40	53.10	52.10	33.00	74.90	46.50
	KNN	55.4455	19.60	**54.30**	55.40	**54.80**	36.10	69.00	49.00
heart-h	J48	**67.6871**	**29.00**	62.20	**67.70**	64.10	43.30	67.30	57.50
	RF	67.0068	26.90	**62.40**	67.00	**64.30**	**43.60**	**81.80**	**66.50**
	DT	57.1429	24.70	58.50	57.10	57.60	32.10	66.20	53.50
	KNN	56.1224	21.30	59.20	56.10	57.60	34.00	69.00	56.90
Indian-liver-patient	J48	68.9537	**51.90**	66.90	69.00	67.60	18.60	67.80	70.70
	RF	**71.3551**	52.70	**68.50**	**71.40**	**69.00**	**22.00**	**74.90**	**77.40**
	DT	67.4099	47.80	67.20	67.40	67.30	19.70	59.80	63.90
	KNN	64.494	45.80	66.50	64.50	65.30	17.90	57.30	62.80
parkinsons	J48	80.5128	**34.40**	80.20	80.50	80.40	46.80	76.90	79.00
	RF	92.8205	16.40	92.70	92.80	92.70	80.10	96.30	**96.90**
	DT	85.641	28.50	85.20	85.60	85.30	59.80	78.60	80.10
	KNN	**96.4103**	4.00	**96.50**	**96.40**	**96.40**	**90.60**	**96.70**	95.30
thyroid	J48	92.093	**10.30**	92.10	92.10	92.10	82.20	90.00	88.20
	RF	94.8837	8.80	94.90	94.90	94.80	88.40	**99.40**	**99.30**
	DT	92.5581	10.20	92.50	92.60	92.50	83.20	91.20	88.60
	KNN	**97.2093**	3.50	**97.20**	**97.20**	**97.20**	**93.70**	96.60	95.80
vertebra-column-2c	J48	81.6129	**27.60**	81.20	81.60	81.20	56.60	83.80	82.30
	RF	**83.2258**	22.10	**83.10**	**83.20**	**83.20**	**61.40**	**92.00**	**92.50**
	DT	79.0323	24.10	79.80	79.00	79.30	53.50	77.50	74.60
	KNN	81.6129	20.30	82.40	81.60	81.90	59.60	80.70	77.50
vertebra-column-3c	J48	82.2581	7.70	82.20	82.30	82.20	74.60	89.50	77.40
	RF	**84.1935**	7.00	**84.20**	**84.20**	**84.10**	**77.30**	**95.90**	**89.50**
	DT	81.2903	8.70	81.30	81.30	81.30	72.60	86.30	73.20



	KNN	78.3871	**11.30**	78.90	78.40	78.50	67.10	83.50	69.70

Based on the results we have obtained from Table 3 all of our datasets are not acceptable. Thus, in this paper, we have proposed model-based machine-learning techniques to classify all our datasets. The model has been presented in Section 3 which consists of several stages. In the first stage, we preprocessed our dataset using data cleaning, filling in missing values, data resampling, and data normalization. Then, the best feature is selected using the PCA algorithm. Finally, we have classified our dataset using the same classifiers which are KNN, J48, RF, and DT. Obtained results of our proposed method are presented in Table 4. Figure 7 presents a plot diagram for biomedical classification.

Figure 7. Biomedical Classification without Preprocessing

Table 4: PIMA and Biomedical Dataset Classification Results (%) Using 10-Fold Cross-validation For Proposed Method

Dataset	Algorithm	Accuracy	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
diabetes	J48	96.2	5.8	96.2	96.2	96.2	91.3	96.7	96.1
	RF	**97.8**	2.5	**97.8**	**97.8**	**97.8**	**95.0**	**99.7**	**99.7**
	DT	97.1	3.0	97.2	97.1	97.1	93.5	97.0	96.0
	KNN	92.2	**9.3**	92.2	92.2	92.2	83.0	96.2	96.3
heart-c	J48	80.5	**7.9**	79.8	80.5	79.6	73.5	92.4	79.8
	RF	**88.8**	4.2	**88.6**	**88.8**	**88.6**	**85.0**	**96.5**	**91.9**
	DT	88.1	4.9	87.9	88.1	87.8	83.7	94.5	86.6
	KNN	87.8	5.5	87.9	87.8	87.5	83.0	96.2	90.1
heart-h	J48	84.7	**12.6**	84.7	84.7	84.7	72.7	87.1	81.7
	RF	**91.8**	10.4	**91.9**	**91.8**	**91.7**	**83.6**	**97.5**	**94.5**
	DT	89.5	9.0	89.8	89.5	89.6	79.9	90.5	85.0
	KNN	90.5	8.9	90.6	90.5	90.3	82.6	96.8	94.0
Indian-liver-patient	J48	85.8	17.4	86.3	85.8	86.0	66.6	88.7	87.5
	RF	**92.5**	11.5	**92.4**	**92.5**	**92.4**	**81.6**	**97.7**	**97.8**
	DT	91.3	11.6	91.3	91.3	91.3	79.0	89.8	88.1
	KNN	92.3	10.2	**92.4**	92.3	92.3	81.4	95.4	95.8
parkinsons	J48	96.4	**11.7**	96.4	96.4	96.3	89.0	95.9	96.8
	RF	**99.0**	3.9	**99.0**	**99.0**	**99.0**	**96.9**	99.1	99.3
	DT	97.4	4.3	97.5	97.4	97.4	92.4	96.6	96.4
	KNN	98.5	5.8	98.5	98.5	98.4	95.3	**99.8**	**99.9**
thyroid	J48	95.8	**8.0**	95.9	95.8	95.8	90.3	94.1	93.2
	RF	93.2	5.7	97.2	**97.2**	**97.2**	**93.6**	**99.6**	**99.4**
	DT	96.3	5.9	96.3	96.3	96.3	91.4	95.2	94.0
	KNN	**97.2**	6.7	**97.3**	**97.2**	**97.1**	**93.6**	99.1	98.9
vertebra-column-2c	J48	91.3	**9.6**	91.4	91.3	91.3	81.3	93.9	92.0
	RF	**94.2**	4.8	**94.5**	**94.2**	**94.2**	**87.9**	**98.6**	**98.6**
	DT	93.2	7.7	93.2	93.2	93.2	85.4	92.8	90.4
	KNN	92.3	7.5	92.5	92.3	92.3	83.6	97.7	97.3
vertebra-column-3c	J48	94.8	**2.4**	95.0	94.8	94.8	92.6	98.1	95.7
	RF	**98.1**	0.8	**98.1**	**98.1**	**98.1**	**97.3**	**99.7**	**99.4**
	DT	97.1	1.4	97.1	97.1	97.1	95.8	97.9	95.1
	KNN	97.4	1.3	97.4	97.4	97.4	96.3	99.4	98.8

Table 5 displays the 4 evaluation matrices for all classification methods that were carried out to the PIMA dataset in order to predict diabetes. These algorithms were used to analyze the data. Table 5 contains a comparative analysis with some recent earlier studies that were found in the research literature. It is known that RF excels in all of the performance metrics and delivers the greatest results for diabetes onset with an accuracy of 97.8%, precision of 97.8%, recall of 97.8%, and F-measure of 97.8%. These figures may be found in the table below.

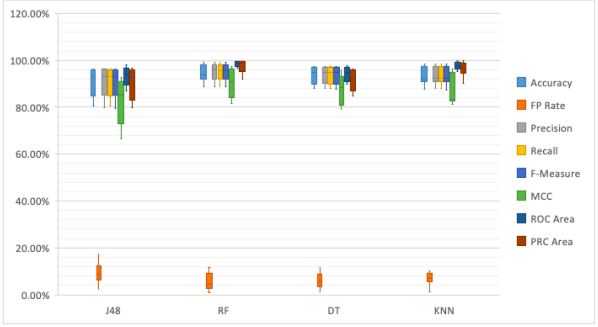


Figure 8: Biomedical Classification with Preprocessing

However, our proposed method has obtained the lowest results using the KNN classifier. Compared to previous works in the study [7, 46, and 47] our work obtained higher results in all matrices with studies [46 and 47] whereas results of the study [7 and 53] obtained good results in accuracy and recall based on DL classifier by obtaining 98.07% accuracy and 98.46% of recall. Figure 8 presents the biomedical classification results after preprocessing stage.

b) Comparison between the proposed method and previous works

In terms of assessing performance, our methodology is superior to that of older work methods, as shown in Table 6. During the stage of preprocessing that comes before the step of data classification, our model makes use of data normalization in order to rescale the data. Our method trains different machine learning techniques. In contrast, some of the earlier research recommends eliminating records with missing data, which generates outcomes that are less than ideal. [54] In the prior approaches, sufficient results were produced by applying fundamental classifiers such as J48, RF, DT, and KNN, in addition to other preprocessing techniques. Despite this, our method is improved by the use of a variety of datasets in the classification process. As it is seen in Table 5, our proposed method obtained better results compared to previous works. [55 and 56]

Table 5: A Comparative Analysis Based PIMA Dataset

Dataset	Algorithm	Accuracy	FP Rate	Precision	Recall
Proposed	J48	96.2	96.2	96.2	96.2
	RF	**97.8**	**97.8**	**97.8**	**97.8**
	DT	97.1	97.1	97.2	97.1
	KNN	92.2	92.2	92.2	92.2
	KNN	97.4	1.3	97.4	97.4
[46]	LR	80.2	82	91	86
	KNN	80.2	84	88	86
	SVM	80.2	82	90	86
	NB	76.56	80	88	84
	DT	77.6	85	81	83
	RF	72.4	77	85	81
[7]	DL	98.07	**98.46**	95.22	96.81
	DT	96.62	95.45	94.02	94.72
	ANN	90.34	83.09	88.05	85.98
	NB	76.33	64.51	59.07	61.67
[47]	J48	75.65	89.92	70.86	79.26
	RF	73.91	79.74	80.79	80.26
	NB	77.83	86.09	81.25	83.60
[49]	RF+PCA	71.44	70.57	-	-
	J48+PCA	71.67	73.81	-	-
	NN+PCA	74.75	73.81	-	-
[50]	KNN	71.72	-	-	-
	DT	66.86	-	-	-
[51]	LR	74.7	-	-	93.51
	SVM	75	-	-	83.67
	RF	73.6	-	-	78.86
	NN	79.8	-	-	77.08
	NB	74.4	-	-	82.27
[52]	RF	95.19	-	-	91.17
	SVM	92.31	-	-	-
	KNN	88.46	-	-	-
	NB	89.42	-	-	-
	DT	91.35	-	-	-
	XGBoost	96.15	-	-	-

Some of the previous works trained their machine learning classifier based on the diabetic dataset without performing preprocessing datasets. Some of the previous works used preprocessing based on data sampling and data normalization which suffer from the fill of missing values. Furthermore, some previous studies performed preprocessing based on data normalization only whereas these methods suffered from other problems-based preprocessing. However, our proposed method attempts to solve most problems of the PIMA dataset based on several steps of preprocessing. First, we cleaned our dataset, then, the mean substitution value is performed to fill in all missing values in our dataset. Next, we resampled our dataset which helps to solve the problem of the unbalanced dataset. Finally, data normalization is applied, based on these steps of preprocessing we found that our dataset has been preprocessed based on obtaining more accurate data for prediction. Furthermore, another step which is feature selection-based PCA is performed to select the most significant features from the data. Based on these steps we found that our method is able to classify the PIMA dataset with high accuracy.

Table 6: A Comparative Analysis Based Biomedical Dataset

Method	classifier	Diabetes	Heart-c	Heart-h	Indian-liver	Parkinsons	Thyroid	Vertabra-2c	Vertabra-3c
Proposed	J48	96.2	80.5	84.7	85.8	96.4	95.8	91.3	94.8
	RF	**97.8**	**88.8**	**91.8**	**92.5**	**99.0**	93.2	**94.2**	**98.1**
	DT	97.1	88.1	89.5	91.3	97.4	96.3	93.2	97.1
	KNN	92.2	87.8	90.5	92.3	98.5	97.2	92.3	97.4
Ref [34]	SVM kernels	-	55.81	67.32	72.38	86.16	**98.16**	85.16	85.16
Ref [35]	LR	-	83.8		-	-	-	-	-
Ref [36]	SVM	-	83.33		-	-	-	-	-
Ref [37]	SVM	-	-	-	77	-	-	-	-
Ref [38]	CNB	-	-	-	71.36	-	-	-	-
Ref [39]	CBC	-	-	-	-		82	-	-
Ref [40]	STN DBS	-	-	-	-	78	-	-	-
Ref [41]	Light GBM	-	-	-	-	93.39	-	-	-
Ref [42]	NB tree	-	-	-	-	-	75	-	-
Ref [43]	RF	-	-	-	-	-	-	**99**	
Ref [44]	LMT	-	-	-	-	-	-	89.73	

V. CONCLUSION

Biomedical data and diabetes mellitus are referred to as just diabetes when referred to in medical contexts. One of a number of recognized metabolic disorders is hyperglycemia (often known as high blood sugar). The diagnosis and management of diabetes are extremely important aspects of patient care in the real world. Diabetes screening is one option that can be used before therapy. In this study, the classification performance of a number of different classifier models is compared. This project's primary objective is to conduct an analysis of the diabetes dataset, making use of support vector machines, decision trees, KNN, and random forest algorithms. The results of this analysis will help in the development of a prediction algorithm and will also contribute to prediction.

REFERENCES

[1] Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besançon, S., ... & Colagiuri, S. (2020). Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*, 162, 108072.

[2] Abdulqadir, H. R., Abdulazeez, A. M., & Zebari, D. A. (2021). Data mining classification techniques for diabetes prediction. *Qubahan Academic Journal*, *1*(2), 125-133.

[3] Arshad, M., Saeed, M., Rahman, A. U., Zebari, D. A., Mohammed, M. A., Al-Waisy, A. S., ... & Thanoon, M. (2022). The Assessment of Medication Effects in Omicron Patients through MADM Approach Based on Distance Measures of Interval-Valued Fuzzy Hypersoft Set. *Bioengineering*, 9(11), 706.

[4] Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1, pp. 727-730, 2016.

[5] Zebari, D. A., Sadiq, S. S., & Sulaiman, D. M. (2022, March). Knee Osteoarthritis Detection Using Deep Feature Based on Convolutional Neural Network. In 2022 International Conference on Computer Science and Software Engineering (CSASE) (pp. 259-264). IEEE.

[6] Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243–6.

[7] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, *19*(1), 391-403.

[8] Ibrahim, D. A., Zebari, D. A., Mohammed, H. J., & Mohammed, M. A. (2022). Effective hybrid deep learning model for COVID-19 patterns identification using CT images. *Expert Systems*, e13010.

[9] Kapoor, N. R., Kumar, A., Kumar, A., Zebari, D. A., Kumar, K., Mohammed, M. A., ... & Albahar, M. A. (2022). Event-Specific Transmission Forecasting of SARS-CoV-2 in a Mixed-Mode Ventilated Office Room Using an ANN. *International Journal of Environmental Research and Public Health,* 19(24), 16862.

[10] Mohammed, H. J., Al-Fahdawi, S., Al-Waisy, A. S., Zebari, D. A., Ibrahim, D. A., Mohammed, M. A., ... & Kim, J. (2022). ReID-DeePNet: A Hybrid Deep Learning System for Person Re-Identification. *Mathematics*, 10(19), 3530.

[11] Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019, April). Machine learning and region growing for breast cancer segmentation. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 88-93). IEEE.

[12] Craven MW, Shavlik JW. Using neural networks for data mining. Futur Gener Comput Syst. 1997;13(2–3):211–29. https://doi.org/10.1016/s0167-739x(97)00022-8.

[13] Radhimeenakshi S. Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural networks. In: 2016 International Conference on Computing for Sustainable Global Development (INDIACom); 2016;3107–11.

[14] Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques, IEEE Access. IEEE. 2019;7: 1365–75

[15] Perveen S, et al. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science. 2016;82:115–21.

[16] Alade OM, Sowunmi OY. Information technology science. 2018;724:14–22.

[17] Putri, NK, Rustam Z and Sarwinda D 2019, Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection *IOP Conf. Ser.: Mater. Sci. and Eng.* **546** 052059

[18] Nurhayati and A N 2014 Implementation of Naive Bayes and K-nearest neighbor algorithm for diagnosis of diabetes mellitus *Proc. of the 13th Int. Conference on Applied Computer and Applied Computation Science* 117-120

[19] Vinoth R *et al* 2014 A Hybrid Text Classification Approach Using KNN and SVM *International Journal of Advance Foundation and Research in Computer (IJAFRC)* **1** 3

[20] Nadira T and Rustam Z 2018 Classification of cancer data using support vector machines with features selection method based on global artificial bee colony *Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Science, AIP Conf. Proc.*

[21] Arfiani, Rustam Z, Pandelaki J, and Siahaan A 2019 Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification *IOP Conf. Ser.: Mater. Sci. and Eng.* 546 052011

[22] Rampisela T V and Rustam Z 2018 Classification of Schizophrenia Data Using Support Vector Machine (SVM) *J. Phys.: Conf. Ser.* 1108 012044

[23] K. Bache and M. Lichman, "UCI Machine Learning Repository," University of California Irvine School of Information, vol. 2008, no. 14/8. p. 0, 2013.

[24] AYDİLEK, İ. B. (2018, September). Examining effects of the support vector machines kernel types on biomedical data classification. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-4). IEEE.

[25] Krstajic, D.; Buturovic, L.J.; Leahy, D.E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminformatics* **2014**, *6*, 10.

[26] Haji, S. H., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., & Zebari, D. A. (2021, July). The Impact of Different Data Mining Classification Techniques in Different Datasets. In *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)* (pp. 1-6). IEEE.

[27] Asaad, R. R., & Ali, R. I. (2019). Back Propagation Neural Network(BPNN) and Sigmoid Activation Function in Multi-Layer Networks. Academic Journal of Nawroz University, 8(4), 216–221. https://doi.org/10.25007/ajnu.v8n4a464

[28] G. Swapna, U. Rajendra Acharya, S. VinithaSree, J. S. Suri, Automated detection of diabetes using higher order spectral features extracted from heart rate signals, Intelligent Data Analysis 17 (2) (2013) 309–326.

[29] Acharya, U. R., Molinari, F., Sree, S. V., Chattopadhyay, S., Ng, K. H., and Suri, J. S., Automated diagnosis of epileptic EEG using entropies. Biomed. Signal Process. Control 7(4):401–408, 2012.

[30] Chicho, B. T., Abdulazeez, A. M., Zeebaree, D. Q., & Zebari, D. A. (2021). Machine learning classifiers-based classification for IRIS recognition. *Qubahan Academic Journal*, *1*(2), 106-118.

[31] Al-Waisy, A. S., Ibrahim, D., Zebari, D. A., Hammadi, S., Mohammed, H., Mohammed, M. A., & Damaševičius, R. (2022). Identifying defective solar cells in electroluminescence images using deep feature representations. PeerJ Computer Science, 8, e992.

[32] Almufti, S., Asaad, R., & Salim, B. (2018). Review on elephant herding optimization algorithm performance in solving optimization problems. International Journal of Engineering & Technology, 7, 6109-6114.

[33] Mohapatra, N., Shreya, K., & Chinmay, A. (2020). Optimization of the random forest algorithm. In *Advances in data science and management* (pp. 201-208). Springer, Singapore.

[34] AYDİLEK, İ. B. (2018, September). Examining effects of the support vector machines kernel types on biomedical data classification. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-4). IEEE.

[35] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*(1), 1-16.

[36] Wang, J. (2021, September). Heart Failure Prediction with Machine Learning: A Comparative Study. In *Journal of Physics: Conference Series* (Vol. 2031, No. 1, p. 012068). IOP Publishing.

[37] Auxilia, L. A. (2018, May). Accuracy prediction using machine learning techniques for Indian patient liver disease. In *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 45-50). IEEE.

[38] A. Anagaw, Y.L. Chang, "A new complement naïve bayesian approach for biomedical data classification", Journal of Ambient Intelligent and Humanized Computing, vol. 10, pp. 3889 - 3897, 2019.

[39] Aversano, L., Bernardi, M. L., Cimitile, M., Iammarino, M., Macchia, P. E., Nettore, I. C., & Verdone, C. (2021). Thyroid disease treatment prediction with machine learning approaches. *Procedia Computer Science*, *192*, 1031-1040.

[40] Habets, J. G., Janssen, M. L., Duits, A. A., Sijben, L. C., Mulders, A. E., De Greef, B., ... & Herff, C. (2020). Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease—proof of principle in a retrospective cohort. *PeerJ*, *8*, e10317.

[41] Nishat, M. M., Hasan, T., Nasrullah, S. M., Faisal, F., Asif, M. A. A. R., & Hoque, M. A. (2021, August). Detection of Parkinson's Disease by Employing Boosting Algorithms. In *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*(pp. 1-7). IEEE.

[42] Turanoglu-Bekar, E., Ulutagay, G., & Kantarcı-Savas, S. (2016). Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms. *Oxford Journal of Intelligent Decision and Data Sciences*, *2*, 13-28.

[43] Reshi, A. A., Ashraf, I., Rustam, F., Shahzad, H. F., Mehmood, A., & Choi, G. S. (2021). Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms. *PeerJ Computer Science*, *7*, e547.

[44] Karabulut, E. M., & Ibrikci, T. (2014). Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing. *Journal of medical systems*, *38*(5), 1-9.

[45] Abdulazeez, A. M., Zeebaree, D. Q., Zebari, D. A., & Hameed, T. H. (2021). Leaf Identification Based on Shape, Color, Texture and Vines Using Probabilistic Neural Network. *Computación y Sistemas*, *25*(3), 617-631.

[46] Patil, V., & Ingle, D. R. (2021, June). Comparative analysis of different ML classification algorithms with diabetes prediction through Pima Indian diabetics dataset. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1-9). IEEE.

[47] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 1-17.

[48] https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e

[49] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, 515.

[50] Theerthagiri, P., & Vidya, J. (2021). Diagnosis and Classification of the Diabetes Using Machine Learning Algorithms.

[51] Saha, P. K., Patwary, N. S., & Ahmed, I. (2019, December). A widespread study of diabetes prediction using several machine learning techniques. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*(pp. 1-5). IEEE.

[52] Samet, S., Laouar, M. R., & Bendib, I. (2021, October). Use of Machine Learning Techniques to Predict Diabetes at an Early Stage. In *2021 International Conference on Networking and Advanced Systems (ICNAS)* (pp. 1-6). IEEE.

[53] Rajab Asaad, R. (2021). Review on Deep Learning and Neural Network Implementation for Emotions Recognition . Qubahan Academic Journal, 1(1), 1–4. https://doi.org/10.48161/qaj.v1n1a25

[54] Taher, K. I., Abdulazeez, A. M., & Zebari, D. A. (2021). Data Mining Classification Algorithms for Analyzing Soil Data. *Asian Journal of Research in Computer Science*, 17-28.

[55] Khalid, L. F., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., & Zebari, D. A. (2021, July). Customer churn prediction in telecommunications industry based on data mining. In *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)* (pp. 1-6). IEEE.

[56] Asaad, R. R., Mustafa, R. F., & Hussien, S. I. (2020). Mortality Statistics and Cause of Death at Duhok City from The Period (2014-2019) Using R Language Data Analytics. Academic Journal of Nawroz University, 9(3), 1–7. https://doi.org/10.25007/ajnu.v9n3a699