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

Data Mining Classification Techniques for Diabetes Prediction

1st *Hindreen Rashid Abdulqadir*
Master Student
Duhok Polytechnic University
Duhok, Iraq
Hindreen.rashid@dpu.edu.krd

2nd *Adnan Mohsin Abdulazeez*
Research Center
Duhok Polytechnic University
Duhok, Iraq
adnan.mohsin@dpu.edu.krd

3rd *Dilovan Assad Zebari*
Research Center
Duhok Polytechnic University
Duhok, Iraq
Diloven.zebari@dpu.edu.krd

https://doi.org/10.48161/qaj.v1n2a55

***Abstract*— Diabetes may be predicted and prevented by exploring critical diabetes characteristics by computational data extraction methods. This study proposed a system biology approach to the pathogenic process to identify essential biomarkers as drug targets. The fact that disease recognition and investigation require many details, data mining plays a critical role in healthcare. This study aims to evaluate the efficiency of the methods used that are based on classification. Besides, the researchers have highlighted the most widely employed techniques and the strategies with the best precision. Many analyses include multiple Machine Learning algorithms for various disease assessments and predictions to improve overall issues. The detection and prediction of diseases is an aspect of classification and prediction. This paper estimates diabetes by its key features and also categorizes the relations between conflicting elements. The recursive random forest removal function provided a significant feature range. Random Forest Classifier investigated the diabetes estimate. RF offers 75,7813 greater precisions than Support Vector Machine (SVM).and may assist medical professionals in making care decisions.**

Keywords*— Data mining, Diabetes prediction, Classification, Random Forest Classifier (RFC), Support Vector Machine (SVM).

I. INTRODUCTION

Nowadays, machine learning (ML) used in every area of computational work where algorithms are designed, and performance is increased [1][2]. In the last years, learning from unbalanced data sets has become a critical problematic in machine learning and is frequently found in several applications such as computer security, Swarm Intelligence [3] [4] , remote sensing [5], biomedicine[6] . data online is massive, and it is growing on a daily basis. It is essential to handle such vast amounts of data and to view the most relevant queries on the user's computer. Since manually analyzing and retrieving relevant data from vast databases is impossible, automatic extraction tools are needed, which enable user-queried data to be retrieved from billions of sites on the internet and relevant knowledge to be discovered. Search engines such as Yahoo, Bing, MSN, and Google are commonly used by users to obtain data from the World Wide Web [7] [8]. Data mining is also used to explore and derive information from data warehouses.

opening up a window of relatively stronger funds, this effective technique increases the sensitivity and/or accuracy of disease identification and diagnosis. The cost of unnecessary and costly diagnostic examinations often declines substantially [2] [3]. For several years' extensive experiments have been carried out in connection with diabetes prediction [4] [5] [6].

Diabetes, also called diabetes mellitus (DM), is a series of metabolic problems that have been detected for a lengthy period from elevated blood glucose levels. Excessive urination, constant thirst and elevated appetite are the symptoms of high glucose. Diabetes can trigger severe health problems such as diabetic ketoacidosis or hyperosmolar hyperglycemic status or may even lead to death if not handled promptly. Which will contribute to a lifespan of coronary distress, stroke in the head, renal dysfunction, foot ulcers and eye problems etc. Diabetes is affected if the pancreas of the body cannot contain enough insulin or if cells and tissues of the body cannot use the insulin provided. [7].

Classification methods are widespread in the medical area for identifying and predicting diseases more accurately [8][9]. Classification is a function for data mining, which assigns groups or groups to objects in a database. In a classification model, for example, applicants may be classified as minimal, medium, or high credit risk[9] [10] [11] [12] [13].Classification has a large number of consumer segmentation uses, including illnesses, business modeling, advertisement, credit analysis [14] [15].

Predictions of diabetes that are not insulin-dependent are used to reduce the risk factor prior to organ damage. The company offers algorithms for the patient dataset (RF) and (SVM) to increase the consistency of attributes. By utilizing machine learning models, too, the authors in the study explained forecast diabetes in the future [16]. And also suggested that they will use more effective data sets to prevent diabetes in this model [17] [18] [19].

The remainder of the paper is structured as follows: Section II includes a related work on the used classification algorithms; Section III contains supplementary details regarding the IRIS datasets; Section IV explains the three approaches used in this study; Section V illustrates the experimental results and discussion; Section VI comparative studies on the mentioned techniques; and Section VII concludes the research work.

II. RELATED WORK

The term "data mining" is a process of assigning individual objects in a database to one or a set of categories or groups. In the phase of classification, the objective is to correctly classify the target class for each instance. This section provides an overview of the most current and useful approaches to classification in various areas of machine learning that have been established by researchers in the last two years. Also, only focuses on, Random Forests(RF), and Support Vector Machine (SVM)as classifiers[10].

Tiwari and Singh [11] discussed the main characteristics estimate diabetes and also categorize the relationship of conflicting characteristics. The recursive removal of features with the random forest was used to make a significant function collection. The detection and prediction of diseases is an aspect of classification and prediction. The calculation of the proposed method states that the Apriority solution is a strong combination of diabetes (BMI) and glucose. The diabetes estimate was tested by XGBoost. Compared with the ANN solution, the XGBoost offers 78.91 percent higher precision and will benefit medical practitioners by decisions about care.

Maniruzzaman et al [12] Explained that Diabetes affected about 422 million people worldwide in 2014. The population is projected to reach 642 million by 2040. The main goal of this study was to develop a tool for predicting diabetic patients who are machine learners (ML). The p-value-and-chances risk factors for diabetes disorders were determined using logistic regression (OR) the study used four classifications, like naive Bayes, to forecast diabetes patients. Precision (ACC) and curve area are used to evaluate the accuracy of these classificatory systems (AUC). Result: The diabetes dataset came from the National Nutrition and Health Survey, which was completed between 2009 and 2012. The data was collected from 6561 interviewees, 657 diabetes monitors, and 5904 diabetes monitors. The combination of LR-based characteristic set and RF-based classification yields an ACC of 94.25 percent and an AUC of 0.95.

Geetha Devasena et al [13] Discussed that Machine-learning algorithms are being used to render possible forecasts. It entails the examination of accessible records. In health care, predictive analytics is mainly used to identify people who are in the early phases of diabetes, hypertension, respiratory failure, or any important lifelong disease. To forecast type 2 diabetes, the suggested system PDD employs data mining algorithms. K-Means Clustering and Random Forest are the data mining algorithms used in the proposed scheme. As opposed to hierarchical clustering and Bayesian network clustering with random-forest prediction, the predictive model PDD produces improved performance in terms of precision.

Rout and Kaur [14] Explained Introduction of early prognosis and diagnostic data mining technology into health care needs greater precision. This paper is intended to review and discuss different study results of data mining methods used in mellitus diabetes. The study found that with the precision of SVM 97.95% the outcome could be called more variables and hybrid disciplines.

Borle et al [15] investigated the difficulties of forecasting potential (Blood Glucose)BG by using a recent T1D dataset of 29,601 entries from 47 different patients, as well as a variety of machine learning algorithms and data preprocessing variants (corresponding to 312 [learner, preprocessed-dataset] combinations). To avoid the harmful long-term consequences of hyperglycemia, patients with Type I Diabetes (T1D) must undergo insulin injections. If we can correctly forecast a patient's predicted blood glucose (BG) levels based on his or her present characteristics, we can create an optimal protocol. They must therefore avoid injecting too much insulin since this will result in (potentially fatal) hypoglycemia. As a result, patients must adhere to a "regimen" that specifies how much insulin to administer at any given moment depending on different calculations. These findings indicate that the standard diabetes diary data may be inadequate to generate accurate BG prediction models; additional data may be required to construct accurate BG prediction models over hourly time intervals.

Vehí et al [16] Explained that the chance of extreme hypoglycemia is the key restricting factor in achieving tight glucose regulation in patients on intense insulin therapy. As a result, hypoglycemia remains the most common safety concern in the management of type 1 diabetes, lowering the standard of life for those who suffer from the condition. This study suggests using four machine learning algorithms to address the problem of diabetes management safety: (1) grammatical evolution for mid-term continuous blood glucose level prediction, (2) support vector machines to predict hypoglycemic events during postprandial periods, (3) artificial neural networks to predict hypoglycemic episodes overnight, and (4) data miniaturization. The plan consists of a mixture of the applied methods' estimation and classification capabilities. The resulting system greatly decreases the number of hypoglycemia episodes, increasing patient protection and giving them more trust in their decision-making.

Maeda-Gutiérrez et al [17] Discussed The aim of this analysis was to discover the factors that predict this complication. A total of 140 topics were included in the dataset, with clinical and preclinical characteristics. The models were assessed using sensitivity, accuracy, the field under the curve (AUC), and the receiving operational characteristic (ROC) curve in a statistical study. The findings show large values obtained by the model using this method, with just three characteristics as predictors accounting for 67 percent of AUC. It is likely to assume that this proposed approach will distinguish patients with DSPN, resulting in a preliminary computer-aided diagnostic tool for the therapeutic field in aiding in the identification of DSPN diagnosis.

Viloria et al [18] Explained that New information and characteristics To evaluate and refine this technique, cCDon diagnosis of DM is needed. Other factors that lead to an accurate diagnosis, such as the concentration of glycosylated hemoglobin A, a biological marker of high significance that often provides an indicator of the type of treatment the patient receives to monitor their disease and health status, may be used to improve precision. In future work, new algorithms, or integrating them with other computational methods like genetic algorithms or particle swarm optimization, may be used to improve the classifier's accuracy and predictability. To make a conclusion, a doctor must examine the outcomes of a patient's examination and equate them to those of other people in similar situations, as well as examine prior decisions.

Ahmad et al [19]Explained that The aim of this paper is to investigate diabetic patient prediction and compare the input features of HbA1c and FPG. By applying five separate machines, we have developed high-performance accuracies, precision, recall, and F1-notes of the model in the dataset by using functional permutations and hierarchy clusters to determine if our data or features are not tied to particular models. In conclusion, we can identify important factors unique to the Saudi population whose management can lead to the control of the disease by the study of the disease utilizing selected features. We also make some of these study suggestions

Albahli [20]Discussed that The hybrid prediction model has an innovative approach which produces far better results than classical computer algorithms. Diabetes is a lifelong condition which may damage different sections of the body for a long time. The objective is to provide a high-precision model for various types of diabetes prediction beginnings. This may aid in early detection of diabetes without excluding findings of missed values. Practices: We are using K-means clustering to apply a noise reduction technology. Afterwards we run the XGB and Random Trees. Our T2ML model achieves higher accuracy than other studies with a precision of 97,53% by 10-fold cross-validation.

III. DATASET

UCI Pima Indian Diabetes Registry is the dataset for the diabetes epidemic. The diabetes dataset contains the attribute name and the attribute description displayed in Table I. The dataset includes 768 tests, each of which has 9 attributes and 1 class with two possibilities such as positive and negative evaluated[21].

Table 1: Decryption of Attributes [21].

S.N	Attributes Description
1	Amount of pregnancy
2	2 hours of oral glucose resistance measure Plasma glucose concentration
3	Blood pressure diastolic (mm Hg)
4	Folding thickness of Triceps tissue (mm)
5	Insulin for 2 hours serum (mm U/m1)
6	Index of body mass (kg/m)^2)
7	Pedigree feature for diabetes
8	Factor Class (tested position or tested negative)
9	Age (year)

IV. METHODOLOGY

A. DATA MINING

Data mining concepts can analyze and classify large volumes of knowledge, group variables with common behaviors, anticipate potential incidents and track and manage patients' privacy[22] [23] .Artificial intelligence subset that utilizes large volumes of data to retrieve substantial data using previously undisclosed pat-predictions [24][25] [26]. Data mining methods have been examined in many health cares trials to forecast the prevalence and characteristics of patients in pandemic conditions. The use of data mining in healthcare companies thus improves service reliability, quality decisions and decreases subjectivity and errors of human beings[27] [28][29].

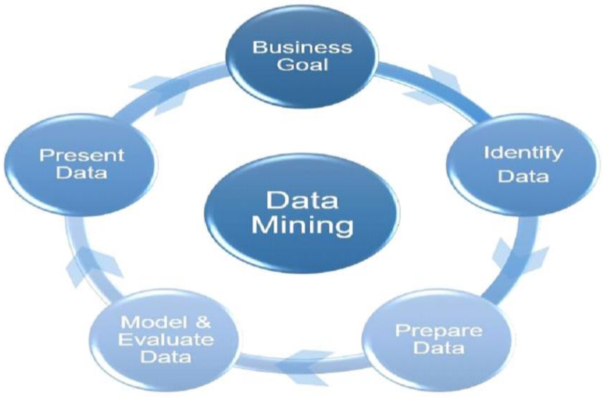


Figure 1:Data mining techniques [30]

B. DIABETES MELLITUS (DM)

In any age category DM is one of the greatest diseases. The more we will deal with the early diagnosis. Machine learning can allow people to make a tentative judgment on DM and can act as a reference for physicians, depending on their daily physical test findings. In DM study, the use of machine learning and data mining methods is a key methodology for utilizing large volumes of accessible knowledge extraction data relevant to diabetes. DM is one of the highest targets of medical study, and ultimately produces a great deal of results, due to the extreme social effects of the individual illness. Computer and data mining approaches in DM are also definitely of considerable importance in terms of diagnosis, control and other clinical management issues. The most important issues for machine learning systems are how to choose the best functionality and classification [31] [32] [33]. Several algorithms have been used lately to forecast diabetes including regular machine learning approaches [34] [35] [36] [37].

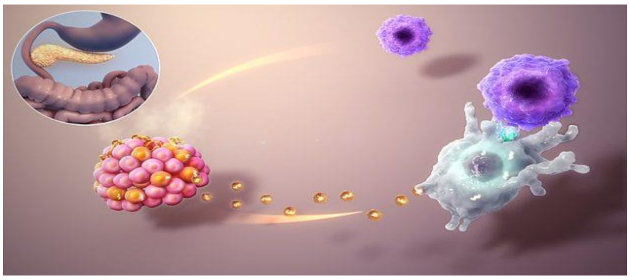


Figure 2:Diabetes Mellitus [38]

C. CLASSIFICATION

The data classification is a two-stage method, consisting of (1) a period in which the actual course of the instance is compared to the intended class and (2) a stage in which a phase in which preparation (or learning) takes place. When the hit rate corresponds to the observer, it is recognized that the classifier is able to distinguish future uncertain class cases. Data classification is a common activity in data mining. There are numerous classifiers used to classify data, including Bayes, function, rule-based, and tree classifiers. Classification aims to predict a specified discrete class variable's value correctly.

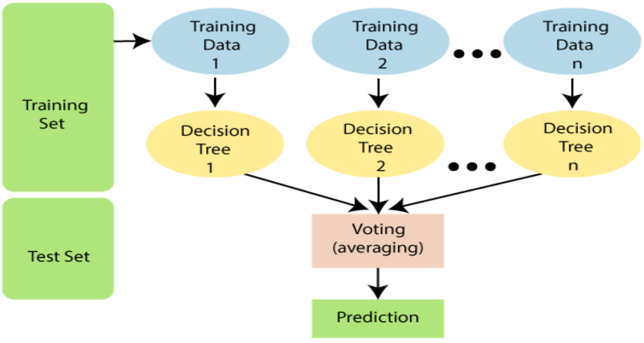


Figure 3: Classification steps of documents [34]

D. RANDOM FOREST CLASSIFIER (RFC)

RFC is an effective decision tree ensemble used for large-scale and multivariate pattern recognition. This ensemble learning is established based on the concept of random subspace recognition. This ensemble learning is established based on the concept of the random subspace method and the stochastic discrimination method of classification. a random forest model becomes a powerful tool to construct an ensemble of classification trees[39]. Successfully applications of RFC have been reported in various studies. RFC is hundreds of Trees of Decision. Each node of the decision tree asks the data and the branches have potential answers to this problem. Random forest combines 100 decision trees. RFC is a supervised form of learning that can deal with classification and regression problems[40].

The algorithm works explain the following steps and Figure 4:

- Choose a random sample from a specific collection of results.
- Set decision trees for each sample, and the outcomes of each decision tree are predicted.
- Select the most forecast outcome for the last forecast [40].

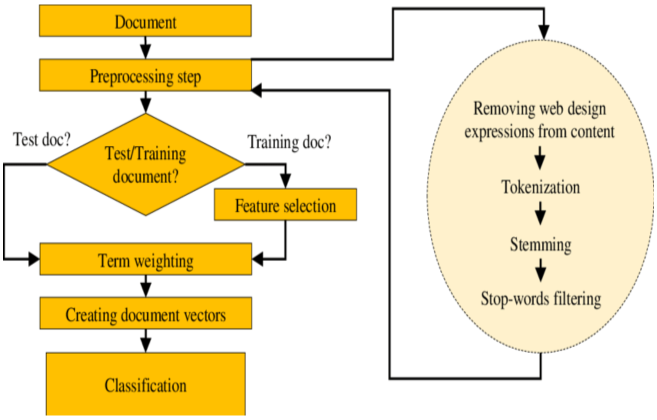


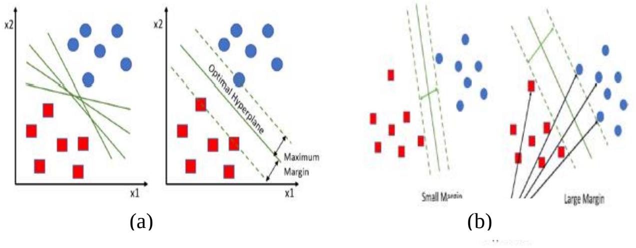
Figure 4:The Random Forest Classification Architecture [40].

Table 2: RF benefits and drawbacks [41]

Benefits	Drawbacks
1) There is better precision. 2) Capable of interacting with huge datasets. 3) It efficiently and easily handles thousands of input variables. 4) Provides detail on critical factors that are not used in the Classifying. 5) Handles missing data while preserving precision. 6) Prototypes are used to include details or meta data on the interaction between multiple variables.	1) One of the more common issues discovered is over-fitting a single data collection, especially in regression tasks. 2) Random Forests struggle with multi-valued and multivalued attributes, in several dimensions. 3) They favor categorical variables of several levels.

E. SUPPORT VECTOR MACHINE (SVM)

The Vector Machine Support (SVM) is a supervised classifier used for regression as well as classification machine learning algorithms. It is mostly used to resolve issues of grouping. SVM aims to categorize data points into a multidimensional space by a sufficient hyperplane. The judgment limits to define data points are hyperplane. The hyperplane classifies the data points with the highest class-hyperplane margin [42]. SVM is a controlled model of computer education. It may be used with a limited collection of data containing fewer outliers. The aim is to split data points and a hyper path. This hyperplane divides space into various fields and contains a category of data for each domain. There are several hyperplanes to be selected to distinguish the two classes of data[40] [43] [51].



(a) (b)

Figure 5: (a): Support Vector Machine with the hyperplane. (b): Support Vector [40]

Table 3: SVM benefits and drawbacks [44]

Benefits	Drawbacks
1) When there is a clear division margin between groups, SVM functions reasonably well. 2) In high dimensional spaces, SVM is more effective. 3) In cases where the number of dimensions exceeds the number of samples, SVM is efficient. 4) SVM is relatively memory efficient	1) When there is a clear division margin between groups, SVM functions reasonably well. 2) In high dimensional spaces, SVM is more efficient. 3) In cases where the number of dimensions exceeds the number of samples, SVM is efficient. 4) SVM is very effective in memory

V. EXPERIMENTAL RESULTS AND DISCUSSION

The results of applying the (RF) and (SVM) algorithms to the diabetes datasets, which are explained in detail in the" TABLE 1", were discovered using a web-based framework developed with Weka. Weka is a software package that includes visualization resources and algorithms for data mining and predictive analytics, as well as graphical user interfaces for quick access to these features. WEKA Engine: WEKA is a compilation of data mining machine learning algorithms. The algorithms may be explicitly implemented to a dataset or from our own Java code named. WEKA includes preprocessed, classified, regressed, clustered, assignment rules, and viewed data resources. It is also suitable for designing new systems in machine learning. We used WEKA as a data mining engine and built a bridge between the framework between Diabetes Expert System and WEKA. [30]. The simulation was performed on a laptop with a Core-i5 processor operating at 2.50 GHz and 4 GB of RAM.

A. *RANDOM FOREST CLASSIFIER (RFC)*

In classification, Cross validation tests of various kinds are used. The researcher evaluated the Random Forest's success using a 10-fold cross validation test in this scenario. The RF is checked on one-fold, while the other folds are used for training. The entire test is replicated five times, with the findings eventually being combined. In all cases of implementation, the results do not change. Only the time spent in the process changes. Where the first implementation process takes longer than in other cases, and the average absorption in all cases is equal to 0.02 seconds. The accuracy in all cases does not change and is equal to 73.2342 % Cross validation is impacted by the time taken for building the model and the accuracy of the process. The execution of the process is repeated more than five times and each time the cross-verification value changes, its value ranges between 10 and 66, so the time elapsed and the accuracy of the process varies each time. The time taken to build the model reaches 0.01 seconds as an average and the accuracy reaches 79.2208 %as an average. Also, the percentage split affects the time it takes to build the model and changes its value from 40 to 90, as the time taken varies and ranges from 0 seconds to. However, the accuracy in all cases tested in the process remains unchanged and remains the same. In addition, when cross validation is set to 10, accuracy by class is detailed in "TABLE 4". In contrast, "TABLE 5" shows the RF efficiency metric. The weighted average accuracy is 75.7813 %, and it takes 0.01 seconds to build a model.

Table 4: Detailed Accuracy of RF by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.836	0.388	0.801	0.836	0.818	0.458	0.820	0.886	tested negative
0.612	0.164	0.667	0.612	0.638	0.458	0.820	0.679	tested positive
0.758	0.310	0.754	0.758	0.755	0.458	0.820	0.81	
Weighted Avg. 2.206	0.862	2.222	2.206	2.211	1.374	2.46	2.375	

Table 5: Performance metric of RF

Attribute	Value
Correctly Classified Instances	75.7813 %
Incorrectly Classified Instances	24.2188 %
Kappa statistic	0.4566
Mean absolute error	0.3106
Root mean squared error	0.4031
Relative absolute error	68.3405 %
Root relative squared error	84.5604 %
Total Number of Instances	768

B. *SUPPORT VECTOR MACHINE (SVM).*

In classification, Cross validation tests of various kinds are used. The researcher evaluated Support vector machine success using a 10-fold cross validation test in this scenario. The SVM is checked on one-fold, while the other folds are used for training. The entire test is replicated five times, with the findings eventually being combined. In all cases of implementation, the results do not change. Only the time spent in the process changes. Where the first implementation process takes longer than in other cases, and the average absorption in all cases the time taken to build the model reaches 0.01 seconds as an average and the accuracy reaches 68.1818 % as an average. Is equal to 0.11 seconds. The accuracy in all cases does not change and is equal to 65.1042 %Cross validation is impacted by the time taken for building the model and the accuracy of the process. The execution of the process is repeated more than five times and each time the cross-verification value changes, its value ranges between 20 and 80, so the time elapsed and the accuracy of the process varies each time. Also, the percentage split affects the time it takes to build the model and changes its value from 40 to 90, as the time taken varies and ranges from 0.06 seconds to However, the accuracy in all cases tested in the process remains unchanged and remains the same. In addition, when cross validation is set to 10, accuracy by class is detailed in "TABLE 6". In contrast, "TABLE 7" shows the SVM efficiency metric. The weighted average accuracy is 65.1042 %, and it takes 0.29 seconds to build a model.

Table 6: Detailed Accuracy of SVM by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	1.000	0.651	1.000	0.789	?	0.500	0.651	tested negative
0.000	0.000	?	0.000	?	?	0.500	0.349	tested positive
0.651	0.651	?	0.651	?	?	0.500	0.546	
Weighted Avg 1.651	1.651	0.651	1.651	0.789	?	1.666	1.546	

Table 7: Performance metric of SVM

Attribute	Value
Correctly Classified Instances	65.1042 %
Incorrectly Classified Instances	34.8958 %
Kappa statistic	0
Mean absolute error	0.349
Root mean squared error	0.5907
Relative absolute error	76.7774 %
Root relative squared error	123.9347 %
Total Number of Instances	768

The researchers evaluated numerous classifiers on diabetes datasets, including Random Forest (RF), and Support Vector Machine (SVM). The results indicate that due to their differences in functionality, the classifiers offer different resolutions on different datasets. The accuracy, error rate, and time taken to build the model were all described in "TABLE 8". In comparison to Random forests and Support Vector Machine (SVM). Algorithm performs exceptionally well. The ultimate conclusion of this paper is that Random Forest has the maximum accuracy, minimum error rate, and takes less time to build the model than other classifiers. Also, the Random forest is graded second in terms of accuracy compared to the decision tree, which is ranked last, even though the time taken to build the model is less than that of the random forest, as seen in the table below, and in the "chart 1". The accuracy was determined using the "Equations (1, 2)" as seen below.

$$\text{Accuracy} = \frac{TP + TN}{P + N} \qquad (1)$$

$$\text{Error Rate} = \frac{FP + FN}{P + N} \qquad (2)$$

The above-mentioned terms in the uncertainty matrix can be discussed in the following details: True Positive (TP) is the product of the positive class being predicted correctly by the model; True negative (TN) results in an accurate projection on the negative class by the model; false positive (FP) is an outcome in which a positive class is mistakenly predicted by the model; false negative (FN) results in the negative class prediction by the model incorrectly; Status Optimistic (P) the number of true positive data cases; status Negative (N) the number of actual negative data cases.

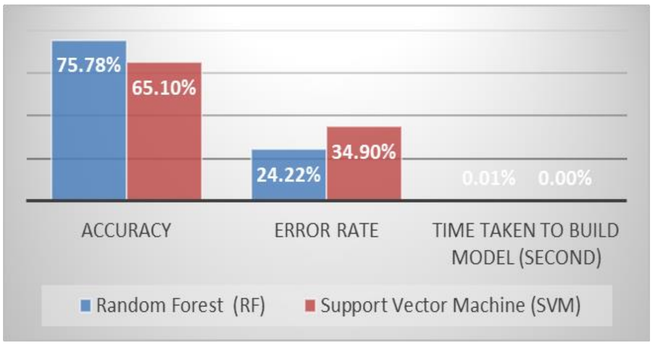


Chart 1: Compares the Accuracy and Error Rate of classifiers

Table 8: Comparison of accuracy, mistake rate and time needed to create Diabetes dataset model.

Classifier	Accuracy	Error Rate	Time taken to build model (second)
Random Forest (RF)	75.7813 %	24.2188 %	0.01%
Support Vector Machine (SVM)	65.1042 %	34.8958 %	0.00%

VI. COMPARATIVE STUDIES

Several classification algorithms based on the classifier used in this study were recorded in related works in this article, illustrating the essential tasks that the researchers with each method tested posed. In this part, the results achieved in this study were compared with the studies that the research has performed in the related work. Analysis [15] used RF on diabetes datasets to increase their efficiency accuracy for 94.25%accuracy for RF, as seen in related work. Compared to this research, the author used RF to check for diabetes but got better results, with an RF accuracy of 75.7813 %.

In the Study [21], many optimization methods were used to classify diabetes datasets. In comparison to other approaches, such as the SVM method had the best precision, which was 99.2, according to the evaluation results. However, in this Study, the employed classifier performed on the same datasets, namely diabetes, while the researcher obtained better performance, which was 65.1042 % obtained by SVM. Although the handprint recognition method was performed using a variety of approaches in datasets. In the Study [17]. The most substantial outcome achieved between them is that it has a 97.95% precision. In comparison, three classifiers are used on diabetes based on this research, with much better results than Study [17], where the l classifier is the SVM 75.7813 % and RF 75.7813 %.

Finally, when contrasting the findings of related work research algorithms to the results of this study, the results of this study tend to be higher, as seen in "TABLE 11". Compared to the other algorithms, RF worked higher than an SVM with 75.7813 % accuracy and an error rate of 24.2188 %. Also, "TABLE 12" outlines the relation of this analysis to the studies analyzed by the researchers in the related work.

Table 9: comparison between this study and related work

Study	Dataset(s)	Classifier	Accuracy
[11]	Pima Indians dataset (PID)	ANN XGboost	ANN 71.35 XGboost 78.91
[12]	(PID)	- (NB) - (RF) - (AB) - (DT)	-NB 86. 67 -RF 89.52 - AB 90.79 -DT 92.54
[13]	K-means clustering algorithm	5-nearest neighbor, and Bayesian network	NN -95.03 %
[14]	-BMI -(PID)	- SVM - dYG	SVM-97.95%
[15]	record set, feature set	- T1D - Gaussian Process Regression	- GPR (48.65 mg/dl).
[16]	1^{41} and dataset 2^{42}	- NCD - (SVC)	- NCD (66%).
[17]	D1 and D2	- RF	RF: 65%.
[18]	(PID)	- (SVM)	-(SVM) (66.25%)
[19]	- HbA1c - FPG	- LG - SVM - DT - RF	-Logistic Regression 80.86 - SVM 82.10 - Decision Tree 74.07 - Random Forest 81.48 -
[20]	(PID)	- LG - NB - M.L.P - KNN - DT	Logistic regression 78.70 Naive Bayes 77.60 Support vector machine 76.95 M.L.P. classifier 74.72 K nearest neighbor 70.83 Decision tree 69.26
This Study	(PID)	- RF - SVM	RF: 75.7813 % SVM: 65.1042 %

VII. Conclusion

In this analysis, we have suggested a system biology approach to the pathogenic pathway to identify essential biomarkers as drug targets and a plan to develop a new multi-molecular drug targeting for diabetes. Diabetes may be anticipated and prevented by analyzing important diabetes characteristics using data extraction techniques. A medical administration sees these motivational and necessary demands for Machine Learning. The principle of Machine Learning quickly pleased health organizations. Diabetes is a growing human disease that requires daily monitoring. To be sure that we find different machine learning algorithms to help before the estimation of the condition. The added model uses an intense gradient technique to increase random forest classification (RFC). The Machine Learning Diabetes library data collection is included in this application. The accuracy of the proposed algorithm improved the support vector machine, contrasting with others (SVM). The precision of our approach adopted is 75,7813.

ACKNOWLEDGEMENT

The researchers would like to indicate that Duhok Polytechnic University supported this study with all of its financial services and motivation.

References

[1] M. J. H. Mughal, “Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, 2018, doi: 10.14569/IJACSA.2018.090630.

[2] J. M. Dennis, “Precision Medicine in Type 2 Diabetes: Using Individualized Prediction Models to Optimize Selection of Treatment,” Diabetes, vol. 69, no. 10, pp. 2075–2085, Oct. 2020, doi: 10.2337/dbi20-0002.

[3] Sulaiman, M. A. (2020). Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. Journal of Soft Computing and Data Mining, 1(2), 11-25.

[4] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,” Procedia Comput. Sci., vol. 82, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[5] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” J. Appl. Sci. Technol. Trends, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[6] A. M. Abdulazeez, D. M. Hajy, D. Q. Zeebaree, and D. A. Zebari, “Robust watermarking scheme based LWT and SVD using artificial bee colony optimization,” Indones. J. Electr. Eng. Comput. Sci., vol. 21, no. 2, pp. 1218–1229, 2021.

[7] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” Procedia Comput. Sci., vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[8] I. Ibrahim and A. Abdulazeez, “The Role of Machine Learning Algorithms for Diagnosing Diseases,” J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 10–19, 2021.

[9] 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.

[10] Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097-203116.

[11] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, “Gene Selection and Classification of Microarray Data Using Convolutional Neural Network,” in 2018 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Oct. 2018, pp. 145–150, doi: 10.1109/ICOASE.2018.8548836.

[12] G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” in 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, Jul. 2013, pp. 1–7, doi: 10.1109/ICCCNT.2013.6726842.

[13] D. Qader Zeebaree, A. Mohsin Abdulazeez, D. Asaad Zebari, H. Haron, and H. Nuzly Abdull Hamed, “Multi-Level Fusion in Ultrasound for Cancer Detection based on Uniform LBP Features,” Comput. Mater. Contin., vol. 66, no. 3, pp. 3363–3382, 2021, doi: 10.32604/cmc.2021.013314.

[14] A. F. Jahwar and A. M. Abdulazeez, “META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING: A REVIEW,” p. 20, 2021.

[15] D. A. Hasan and A. M. Abdulazeez, “A Modified Convolutional Neural Networks Model for Medical Image Segmentation,” p. 12, 2020.

[16] N. M. Abdulkareem and A. M. Abdulazeez, “Machine Learning Classification Based on Radom Forest Algorithm: A Review,” Int. J. Sci. Bus., vol. 5, no. 2, pp. 128–142, 2021.

[17] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, “Machine Learning Supervised Algorithms of Gene Selection: A Review,” Mach. Learn., vol. 62, no. 03, 2020.

[18] S. Dahiya, R. Tyagi, and N. Gaba, “Comparison of ML classifiers for Image Data,” EasyChair, 2020.

[19] P. M. S. Sai, G. Anuradha, and V. P. kumar, “Survey on Type 2 Diabetes Prediction Using Machine Learning,” in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, Mar. 2020, pp. 770–775, doi: 10.1109/ICCMC48092.2020.ICCMC-000143.

[20] P. Tiwari and V. Singh, “Diabetes disease prediction using significant attribute selection and classification approach,” J. Phys. Conf. Ser., vol. 1714, p. 012013, Jan. 2021, doi: 10.1088/1742-6596/1714/1/012013.

[21] Md. Maniruzzaman, Md. J. Rahman, B. Ahammed, and Md. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” Health Inf. Sci. Syst., vol. 8, no. 1, p. 7, Dec. 2020, doi: 10.1007/s13755-019-0095-z.

[22] M. S. Geetha Devasena, R. Kingsy Grace, and G. Gopu, “PDD: Predictive Diabetes Diagnosis using Datamining Algorithms,” in 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, Jan. 2020, pp. 1–4, doi: 10.1109/ICCCI48352.2020.9104108.

[23] M. Rout and A. Kaur, “Prediction of Diabetes Risk based on Machine Learning Techniques,” in 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, Jun. 2020, pp. 246–251, doi: 10.1109/ICIEM48762.2020.9160276.

[24] N. C. Borle, E. A. Ryan, and R. Greiner, “The challenge of predicting blood glucose concentration changes in patients with type I diabetes,” Health Informatics J., vol. 27, no. 1, p. 146045822097758, Jan. 2021, doi: 10.1177/1460458220977584.

[25] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi, “Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning,” Health Informatics J., vol. 26, no. 1, pp. 703–718, Mar. 2020, doi: 10.1177/1460458219850682.

[26] V. Maeda-Gutiérrez et al., “Distal Symmetric Polyneuropathy Identification in Type 2 Diabetes Subjects: A Random Forest Approach,” Healthcare, vol. 9, no. 2, p. 138, Feb. 2021, doi: 10.3390/healthcare9020138.

[27] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, “Diabetes Diagnostic Prediction Using Vector Support Machines,” Procedia Comput. Sci., vol. 170, pp. 376–381, 2020, doi: 10.1016/j.procs.2020.03.065.

[28] H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, and A. Alhumam, “Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning,” Appl. Sci., vol. 11, no. 3, p. 1173, Jan. 2021, doi: 10.3390/app11031173.

[29] S. Albahli, “Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection,” J. Med. Imaging Health Inform., vol. 10, no. 5, pp. 1069–1075, May 2020, doi: 10.1166/jmihi.2020.3000.

[30] S. Joshi and S. PriyankaShetty, “Performance analysis of different classification methods in data mining for diabetes dataset using WEKA tool,” Int. J. Recent Innov. Trends Comput. Commun., vol. 3, no. 3, pp. 1168–1173, 2015.

[31] A. U. Haq et al., “Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data,” Sensors, vol. 20, no. 9, p. 2649, May 2020, doi: 10.3390/s20092649.

[32] V. Kumar, B. K. Mishra, D. N. H. Thanh, and A. Verma, “Prediction of Malignant & Benign Breast Cancer: A Data Mining Approach in Healthcare Applications,” p. 8.

[33] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, “Machine learning and Region Growing for Breast Cancer Segmentation,” in 2019 International Conference on Advanced Science and Engineering (ICOASE), Zakho - Duhok, Iraq, Apr. 2019, pp. 88–93, doi: 10.1109/ICOASE.2019.8723832.

[34] A. S. Eesa, A. M. A. Brifcani, and Z. Orman, “A New Tool for Global Optimization Problems- Cuttlefish Algorithm,” vol. 8, no. 9, p. 6, 2014.

[35] M. L. Kolling et al., “Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development,” Int. J. Environ. Res. Public. Health, vol. 18, no. 6, p. 3099, Mar. 2021, doi: 10.3390/ijerph18063099.

[36] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[37] D. A. Zebari, H. Haron, D. Q. Zeebaree, and A. M. Zain, “A Simultaneous Approach for Compression and Encryption Techniques Using Deoxyribonucleic Acid,” in 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Island of Ulkulhas, Maldives, Aug. 2019, pp. 1–6, doi: 10.1109/SKIMA47702.2019.8982392.

[38] “DATA MINING TECHNIQUES. What is data mining? | by Tanmay Terkhedkar | Medium.” https://medium.com/@tanmayct/data-mining-techniques-24d01a8fb71e (accessed Apr. 18, 2021).

[39] M. Li, X. Fu, and D. Li, “Diabetes Prediction Based on XGBoost Algorithm,” IOP Conf. Ser. Mater. Sci. Eng., vol. 768, p. 072093, Mar. 2020, doi: 10.1088/1757-899X/768/7/072093.

[40] N. Singh and P. Singh, “Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus,” Biocybern. Biomed. Eng., vol. 40, no. 1, pp. 1–22, Jan. 2020, doi: 10.1016/j.bbe.2019.10.001.

[41] J. Abdollahi and B. Nouri-Moghaddam, “Hybrid stacked ensemble combined with genetic algorithms for Prediction of Diabetes,” p. 12.

[42] M. M. Hatmal et al., “Artificial Neural Networks Model for Predicting Type 2 Diabetes Mellitus Based on VDR Gene FokI Polymorphism, Lipid Profile and Demographic Data,” Biology, vol. 9, no. 8, p. 222, Aug. 2020, doi: 10.3390/biology9080222.

[43] M. U. Emon, R. Zannat, T. Khatun, M. Rahman, M. S. Keya, and Ohidujjaman, “Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models,” in 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, Jan. 2021, pp. 1048–1052, doi: 10.1109/ICICT50816.2021.9358612.

[44] S. Chang, J.-Y. Chen, Y.-J. Chuang, and B.-S. Chen, “Systems Approach to Pathogenic Mechanism of Type 2 Diabetes and Drug Discovery Design Based on Deep Learning and Drug Design Specifications,” Int. J. Mol. Sci., vol. 22, no. 1, p. 166, Dec. 2020, doi: 10.3390/ijms22010166.

[45] O. Diouri et al., “Hypoglycaemia detection and prediction techniques: A systematic review on the latest developments,” Diabetes Metab. Res. Rev., Mar. 2021, doi: 10.1002/dmrr.3449.

[46] “Diabetes Mellitus — Types, Complications and Treatment | Medical Library,” The Lecturio Online Medical Library, Oct. 12, 2015. https://www.lecturio.com/magazine/diabetes-mellitus/ (accessed Apr. 18, 2021).

[47] V.-H. Dang, N.-D. Hoang, L.-M.-D. Nguyen, D. T. Bui, and P. Samui, “A novel GIS-based random forest machine algorithm for the spatial prediction of shallow landslide susceptibility,” Forests, vol. 11, no. 1, p. 118, 2020.

[48] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, “A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic,” IEEE Access, vol. 9, pp. 7869–7884, 2021, doi: 10.1109/ACCESS.2020.3047942.

[49] D. Devetyarov and I. Nouretdinov, “Prediction with confidence based on a random forest classifier,” in IFIP International Conference on Artificial Intelligence Applications and Innovations, 2010, pp. 37–44.

[50] D. Anguita, A. Ghio, N. Greco, L. Oneto, and S. Ridella, “Model selection for support vector machines: Advantages and disadvantages of the machine learning theory,” 2010, pp. 1–8.

[51] Saeed, J. N., & Abdulazeez, A. M. (2021). Facial Beauty Prediction and Analysis Based on Deep Convolutional Neural Network: A Review. Journal of Soft Computing and Data Mining, 2(1), 1-12.

[52] Abdulkareem, N. M., Abdulazeez, A. M., Zeebaree, D. Q., & Hasan, D. A. (2021). COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms. Qubahan Academic Journal, 1(2), 100-105.

[53] Muhammad, M., Zeebaree, D., Abdulazeez, A. M., Saeed, J., & Zebari, D. A. (2020). A Review on Region of Interest Segmentation Based on Clustering Techniques for Breast Cancer Ultrasound Images. J. Appl. Sci. Technol. Trends, 1(3), 78-91.

[54] Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine Learning Applications based on SVM Classification A Review. Qubahan Academic Journal, 1(2), 81-90.

[55] Zebari, D. A., Zeebaree, D. Q., Saeed, J. N., Zebari, N. A., & Adel, A. Z. (2020). Image steganography based on swarm intelligence algorithms: A survey. people, 7(8), 9.

[56] Taha, M. S., Rahim, M. S. M., Hashim, M. M., & Khalid, H. N. (2020). Information Hiding: A Tools for Securing Biometric Information. Technology Reports of Kansai University, 62(04), 1383-1394.

[57] Adeen, N., Abdulazeez, M., & Zeebaree, D. Systematic Review of Unsupervised Genomic Clustering Algorithms Techniques for High Dimensional Datasets.

[58] Abdulazeez, A. M., & Faizi, F. S. (2021). Vision-Based Mobile Robot Controllers: A Scientific Review. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(6), 1563-1580.