iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.45081

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Diabetes Prediction using Machine Learning: A Review

Anushka Awasthi[1], Ishwar Gangwal[2], Mihir Jain[3]
[1]*Department of Electronics and Instrumentation Engineering, IET, DAVV*
[2, 3]*Department of Electronics and Telecommunications Engineering, IET, DAVV*

Abstract: Prediction of diabetes using machine learning algorithms has been thoroughly studied by several researchers in the past. Finding, critically assessing, and combining the data of all pertinent, high-quality individual research is what prompted us to conduct this assessment of multiple diabetes prediction models. Analysis of several writers' expertise on diabetes prediction systems is presented in this publication. This study on diabetes prediction models aimed to discover the best strategies for selecting and synthesising the many studies of high quality. The majority of medical data is nonlinear, correlation-structured, and complicated, making it difficult to analyse. The use of machine learning-based techniques in healthcare and medical imagery has been ruled out.

I. INTRODUCTION

As part of this research, this paper has looked at a number of machine learning-based approaches for making diabetes predictions and then compared them.

The study's goals are as follows:

1) In order to broaden our knowledge of the numerous diabetes prognostication models.
2) To determine the correctness of the current machine learning models and discuss them.
3) To go through the numerous characteristics needed to predict diabetes.
4) An evaluation of several diabetes prediction models is described in this paper.

A thorough review of all relevant, high-quality individual research and a rigorous analysis of the resulting data prompted us to look at several diabetes predictions models. The researcher has read a lot of papers on diabetes prediction models to help fuel our enthusiasm for this evaluation process.

A. Diabetes

Diabetes is a condition that is brought on by having an abnormally high level of blood glucose in the body. Human bodies are in constant need of power, and sugar is one of the primary sources of vitality that is used in the construction of our muscles and other tissues. In individuals, the primary reasons of type 2 diabetes are often an unhealthy habit combined with a lack of physical activity. Diabetes is a condition that is brought on by an abnormally high level of glucose in the bloodstream. Diabetes occurs when the pancreas is failing to turn the meal into insulin; as a result, sugar is not taken into the body, leading to the condition. Diabetes may cause problems in a variety of body systems, including the kidneys, eyes, neurological system, arteries, and so on. There are three different forms of diabetes. The first kind of diabetes is known as juvenile diabetes (Sun & Zhang, 2019) which primarily affects youngsters and damages the cells in the pancreatic that are responsible for insulin production. The second form of diabetes is called type 2, and it is often diagnosed in people over the age of 40 who do not get enough exercise and have bad lifestyles. Diabetes is a form of illness that cannot be cured but may be managed well via a healthy diet, exercising regularly, and the use of appropriate medication. Diabetes cannot be reversed (Sun & Zhang, 2019).

People with type 2 diabetes (Malik, et al., 2020) do not receive insulin injections at regular intervals, which is why this form of diabetes is often referred to as insulin-independent diabetes. Patients with type 1 diabetes, on the other side, receive insulin injections at periodic intervals, which is why this form of diabetes is often referred to as insulin-dependent diabetes (Malik, et al., 2020). The third kind of diabetes (Han, et al., 2020) is called gestational diabetes, and it is caused by the shift in hormone levels that happens during childbirth. In several cases, gestational diabetes goes away once the child is born. Prediabetes is a situation in

wherein patient's blood sugar levels are borderline for diabetes, however this situation may be corrected with the aid of physical activity and a healthier diet. Prediabetes is the only diagnosis in which blood sugar levels are borderline for diabetes.

B. Need of Machine Learning

The field of artificial intelligence known as machine learning is concerned with the process by which computers attempt to forecast future events using historical data and the information they already possess. There are two distinct forms of machine learning. The first kind of learning is called supervised learning, and in this type of learning, the data itself serve as the instructor, and the system is constructed based on the dataset. The second kind of learning is called unsupervised learning, and it involves the data teaching itself by identifying certain patterns within the dataset and then categorising those patterns. Over the last several years, a large number of writers have reported and discussed their research on diabetes prediction by utilising machine learning algorithms.

C. Machine Learning algorithms

The research that has been done on machine learning has resulted in the development of multiple data mining methods. These algorithms may be directly applied to a dataset in order to create some predictions or to derive important inferences and conclusions from such a dataset. Immediate use of these algorithms is possible. Decision tree, Naive Bayes, k-means, neural network, and other similar algorithms are examples of prominent data mining techniques. In the part that comes after this one, we will talk about them.

1) Decision Tree A decision tree is a kind of decision support system that utilises a tree-like structure or representation of actions and their potential repercussions, which may include the results of chance events as well as utility considerations. It is one of the methods in which an algorithm may be shown. In the field of operational research, decision trees are often used, particularly in the process of decision analysis, to assist with the identification of a strategy that would most likely result in the achievement of the objective. In the field of machine learning, it is also a widely used tool. The process of mapping first from tree's parent node to each of its leaf nodes individually may quickly and efficiently convert a decision tree to a collection of rules. When everything is said and done, reasonable judgments may be obtained by adhering to these criteria.

2) C4.5 It takes the shape of a decision tree and functions as a classifier. It is a sort of learning known as supervised learning that makes use of information acquisition and pruning in order to get better outcomes. It is quite famous, and the results may be understood in a straightforward manner.

3) K-means Algorithm – K-means organizes the items in a set into k groups, with the goal of making the elements inside each group more comparable to one another. The k-means clustering method, in addition to allowing the user to choose the quantity of clusters, "learns" the groupings on its own despite the user providing any information regarding which cluster each individual observation must be placed in. Because of this, the k-means algorithm is often referred to as a semi-supervised learning approach. The K-means algorithm performs very well when used to huge datasets.

4) The ID3 method, which is also known as the Quinlan86 algorithm, is a decision tree construction technique that decides the categorization of objects by examining the numbers of the attributes. It constructs the tree in a top-down manner, beginning with a set of entities and the definition of their characteristics as the building blocks. At each tree node, a characteristic is examined, and the results of those examinations are used to define the partitioning criteria for the entity at that point. This method is repeated in a recursive manner until the set contained inside a particular subtree is uniform with regard to the classification criterion. After then, it transforms into a leaf node. At every node, the amount of information gained is increased while the amount of entropy is decreased. To put it another way, the quality that tests how well the candidate set can be divided into subgroups with similar characteristics is evaluated.

5) Support Vector Machine (SVM) - an abbreviation for It is a kind of learning called supervised learning, and it divides data into two categories using a hyper plane as the dividing line. The Support Vector Machine (SVM) accomplishes the same goal as C4.5, with the exception that it does not make any use of Decision Trees. In order to reduce the likelihood of an incorrect classification being made, the support vector machine makes an effort to increase the margin, which is the distance across the hyper plane and the 2 data points that are closest to it from each class. Scikit-learn, MATLAB, and LIBSVM are examples of well-known software packages that may be used to create support vector machines.

6) The Naive Bayes (NB) method is a straightforward approach to the construction of classifiers. It is a Bayesian probabilistic classifier that uses Bayes' theorem as its foundation. Given the class variable, each and every Naive Bayes classifier operates on the assumption that the value of a single feature does not rely in any way on the value of any additional feature. The following statement illustrates Bayes theorem: Here X is the data tuple and C is the class, P(C|X) is calculated by multiplying P(X|C) by P(C)/P(X). This ensures that P(X) remains the same for all classes. Despite the fact that it operates on the assumption that

feature values are critically independent, which is not a realistic assumption, it performs very well on huge datasets when this requirement is assumed and is true.

7) A Network Made of Artificial Neurons (ANN) A computer model that replicates the architecture and functionality of biological neurons is referred to as an artificial neural network, or ANN for short. Since a neural network adapts or trains in some manner depending on the outputs and inputs, for that specific stage and subsequently for each phase, data that travels thru the network has an effect on the architecture of the ANN. ANNs are recognised as nonlinear statistical data modelling software because of their ability to simulate the intricate interactions that exist across inputs and outputs or to identify trends. ANNs are built using layers that are linked to one another. The artificial neural networks that are being used to improve current data analytics tools are very straightforward mathematical models.
8) Classification or Regression Trees technique is what is meant by the acronym CART. A categorized outcome variable is often utilized in classification trees, and the purpose of the tree is to determine which "class" the categorized target variable most likely belongs to within the larger set of classes. When using regression techniques, the target parameter is assumed to be continuous, and a tree is utilised to make predictions about the target variable's quantity. The CART algorithm is organised as a succession of questions, the responses to which indicate what will be the following issue if there would be any queries. The answers to these inquiries are presented in the form of a tree, each branch of which culminates in a single node to indicate that there are no further questions to be answered.
9) Random Forests are a kind of ensemble learning approach that may also be considered a type of closest neighbour predictor for the purposes of classification and regression procedures. During the training phase, it builds a number of decision tree and produces an output that would be the class that is most common among the classes produced by the individual trees. It also seeks to mitigate the concerns of high variation and large bias by aggregating to establish a natural equilibrium across the two extremes. There are strong package implementations available for this approach in both R and Python.
10) Regression is a statistics concept that is used to evaluate the strength of a connection among one dependent variable (often represented by Y) and a number of other varying variables. This is done by comparing the two sets of data and analysing the results. Regression may be broken down into two primary categories: linear regression and multiple linear regression. In addition, there are a number of non-linear regression techniques that may be utilised for data analysis of a more complex kind.

II. RELATED WORK

Patients may get assistance in resuming their normal routines of life via the provision of individualised services in a variety of medical specialties offered by healthcare systems. The condition known as diabetes mellitus ranks among the most critical and severe challenges faced by the medical community. In the current set of realistic conditions, classification is one of the most important decision-making approaches that may be used. The major objective is to classify the data as either being related to diabetes or not being related to diabetes, as well as improve the classification accuracy (Saxena, et al., 2022). When it comes to the diagnosis of diabetes, machine learning is primarily focused on recognising patterns within the diabetes dataset that would be provided. In recent years, machine learning has emerged as the most reliable and helpful innovation in the field of medicine, and this trend is expected to continue in the foreseeable future. With the use of machine learning classifiers, the primary objective of this work is to categorise diabetes patients into different kinds depending on the information they provide about themselves and their clinical conditions. This section includes an overview of the works that were proposed by various researchers during the course of the previous ten years. It is helpful to detect the inadequacies of recommended works in the area of machine learning classifiers for diabetic patients' treatment regimens. The identification of diabetes is becoming an increasingly important topic of research (Saxena, et al., 2022).

Several other deep learning approaches and classification methods, including artificial neural networks, decision trees, random forests, and support vector machines, have been described in (Sun & Zhang, 2019) work. In order to classify diabetes-related data, (Qawqzeh, et al., 2020) used a classification strategy based on logistic regression. There are 459 patients included in the training data, and 128 patients are included in the testing data. Utilizing logistic regression, the authors were successful in achieving a classification accuracy of 92 percent. The fact that the model was not compared to any of the other diabetes predictive models and, as a result, was unable to be verified was the model's most significant shortcoming. One half of the dataset was used to train the algorithm while the other half was used to test it. (Qawqzeh, et al., 2020). In order to make a forecast of diabetes, the naive Bayes and support vector machine methods were combined in the framework that was presented. The suggested model was verified on this dataset after the dataset was obtained from three separate places. The dataset had a total of 402 individuals and contained eight different features, one of which was the presence of type 2 diabetes in 80 of the patients (Tafa, et al., 2015). The ensemble of naive

Bayes and SVM has accomplished an accuracy of 97.6 percent, which is significantly higher than the accuracy obtained by either of the algorithms when they were run individually on the dataset, with naive Bayes accomplishing an accuracy of 94.52 percent and support vector machine attaining 95.52 percent respectively. The authors have not stated any pre-processing techniques in order to delete any undesirable inputs from the dataset.

(Karan, et al., 2012) presented a novel approach for diagnosing diabetes by constructing a distributed end-to-end three-level inescapable healthcare system framework using artificial neural network (ANN) computation. This allowed them to show the new method. Sensors and other wearable technology are utilized to measure vital signs and other indications on the human system at the most fundamental level. At the second level, client-side devices like personal digital assistants (PDAs) and personal computers (PCs) act as an arbitrator and mediator across the primary level and the final tier. Customers get assistance with social welfare procedures and database activities from powerful desktop servers, which are part of the third level's culmination (Karan, et al., 2012). In order to identify disorders at both the following and subsequent stages, techniques of an artificial neural network are performed. The client and server paradigm are depending on the calculations of artificial neural networks. Using the idea of sickness as the basis for computations and system communications on both the client and the server sides is how this strategy develops both of those areas.

On the Pima Indians Diabetes Collection, (Sisodia & Sisodia, 2018) used the Naive Bayes, decision trees, and SVM learning methods. The Naive Bayes classifier obtained the highest level of accuracy in its ability to forecast diabetes. Sisodia used a method known as tenfold cross-validation, which consisted of dividing the dataset into 10 equal portions and then using nine of those parts for training purposes while using the tenth part for assessment purposes. Precision, accuracy, recall, and area under the curve were always the evaluation measures that were used to forecast diabetes. (Hussain & Naaz, 2021) provided an evaluation of a number of different machine learning techniques. Within this review, the accuracy of random forest, Naive Bayes, and NN was examined and contrasted. The Matthews correlation coefficient was utilised by the authors in order to carry out the evaluation of these machine learning techniques. The research conducted by (Kumari, et al., 2021) on the Pima Indians Diabetes Dataset included the application of Naive Bayes, RF, and LR. The researchers then contrasted these three methods to an ensemble method and found that the ensemble strategy provided the most accurate results for the model.

Deep learning, often known as a neural network, is a multi-layered network that uses feed forward. (Olaniyi & Adnan, 2014) used this kind of network in their work. The technique was applied to the Pima Indians Diabetes Dataset by the researchers; the dataset was then split in such a manner that 500 entries were utilized for training reasons and 268 entries were utilized for testing reasons. Before any kind of pre-processing activities could be carried out, the dataset was first normalised in order to establish numerical stability. By dividing each characteristic by its associated amplitude, the values of the dataset were normalised such that they all fell within the range of 0 to 1, which was the goal of the normalisation process. The authors were able to attain an accuracy rate of 82 percent with their predictions. SVM and Naive Bayes techniques were used by (Gupta, et al., 2021) in their study to categorise the diabetes dataset. The authors trained and tested their model using a k-fold cross-validation, and when they utilised both classification methods to their data, the SVM classification performed much better than that of the Naive Bayes technique.

(Kandhasamy & Balamurali, 2015) used a few different machine learning algorithms to make a prediction of diabetes using a dataset that has been obtained from the UCI repository. These algorithms included random forest, J48, k-nearest neighbours, and SVM. Both before and afterwards pre-processing the dataset, the authors used the aforementioned classifier to analyse it. The precompiled data was used in the second attempt. There was no discussion of pre-processing procedures; all that was mentioned was the notion that the database contained some noise that was eliminated. The authors have assessed the accuracy, sensitivity, and applicability of their prediction using those three criteria. The decision tree obtained the maximum accuracy of 73.82 percent when the dataset was not pre-processed, whereas the random forest earned the best accuracy of 100 percent when the information was pre-processed.

(Choubey, et al., 2020) used two feature selection approaches called PCA and linear discriminant evaluation to extract relevant characteristics from the Pima Indians Diabetes Dataset. Linear discriminant analysis and Principal component analysis are both types of factor analysis. In addition to that, a comparative study of the strategy for selecting attributes was included in the paper. For the aim of classification, a select group of machine learning methods, including radially foundation kernel, KNN, and AdaBoost, were also used to the dataset in question. The dataset obtained from the Canadian primary healthcare sentinel monitoring network was used in the research carried out by (Perveen, et al., 2016). The parameters that are included in the dataset include sex, BMI, fasting blood sugar, triglycerides, and systolic and diastolic heart rate, respectively. The authors of the study utilised decision trees, bootstraps, and adaptable boosting as their classifiers of choice.

Utilizing machine learning methods, (Gujral, 2017) published a survey on the key phases of diagnosing type 2 diabetes. The survey also identified frequently occurring problems related with diabetic retinopathy and nephropathy.

Quite a few different approaches to machine learning have already been looked at and researched, some of which include artificial neural networks, essential components, decision trees, hereditary computing, and fuzzy logic. The Pima Indians Diabetes Dataset serves as the informational index for the vast bulk of the relevant body of research, which may be found here. It is critical to get an accurate diagnosis of diabetes in the initial stages so that life-threatening complications associated with the disease may be mitigated. Based on the findings (Gujral, 2017) of the Writing Survey of Diabetes Assumptions, it is clear that a solitary approach to diagnosing diabetes is not a very sophisticated way of diagnosing diabetes at an early stage. Combining many types of classifiers, including SVM, principal component analysis, and evolutionary algorithms, together with ANN, yields the best possible results.

(Mamuda & Sathasivam, 2017) used supervised machine learning classifiers such as scaled conjugate gradient, Levenberg–Marquardt, and Bayesian regulation. All of these methods fall under the category of "supervised machine learning." After separating the data into testing and training batches, the Levenberg–Marquardt algorithm demonstrated the highest level of accuracy. (Malik, et al., 2020) worked with a regionally accessible dataset that was acquired from a facility in Germany. They implemented decision trees, KNN, and random forest on top of this locally accessible dataset. The identification of diabetes was accomplished by (Soltani & Jafarian, 2016) through the utilisation of a probabilistic neural network. The Pima Indians Diabetes Dataset was split into two parts: a training dataset consisting of 90 percent of the total, and a testing dataset consisting of the remaining 10 percent. The accuracy that was reached for the training set was 89.56 percent, while the precision for the test dataset was 81.49 percent.

Both (Tigga & Garg, 2020) contributed to the Pima Indians Diabetes Dataset. Blood glucose levels, the quantity of pregnancies, and BMI were shown to be three of the most important parameters collected from the information. RStudio was used to make a prediction about the accuracy utilizing logistic regression, and the result was that the accuracy reached was 75.32 percent. The Pima Indians Diabetes Dataset was analysed using the Naive Bayes, and random forest classification methods by (Yuvaraj & SriPreethaa, 2017). In conjunction, the information gain attribute selection approach was used in additament to machine learning algorithms in order to retrieve the key features. Furthermore, eight features were utilised as opposed to thirteen characteristics as a result of this change. 30 percent of the dataset was employed for testing objectives, and the authors demonstrated that a random forest algorithm achieves a maximum efficiency of 94 percent of the time.

(Rashid, et al., 2016) constructed diabetic mellitus support systems, which operate automatically by using classification algorithms, taking into account various versions of the aforementioned concerns. Also reflecting the capabilities of medical professionals who are definite that there is a substantial association between the negative effects of particular chronic diseases and the frequency of glucose production in the blood. It's possible that the implications of this research transcend beyond just classifying people who have diabetes into different groups. The primary commitments are as follows, given this arrangement: It takes use of a few free variables here and there.

(Negi & Jaiswal, 2016) developed their own unique dataset, which has 102538 items and 49 descriptors altogether. In all, this dataset included around 64419 diabetic individuals, whereas the remaining patients did not have diabetes. Using a pre-processing approach, we were able to fill in the values that were absent, and normality test were converted into numerical data. In order to choose the pertinent characteristics from the dataset, the wrapper element selection approach and the ranking method were both used. To further improve accuracy, an aggregation of a few different classifiers was applied. The improved accuracy was 72%.

The Bayes net, Hoeffding tree, JRip, and multilayer perceptron were some of the models that (Mercaldo, et al., 2017) used to analyse the Pima Indians Diabetes Dataset. In order to choose the relevant features that would lead to an improvement in the efficiency of classifiers, researchers turned to both the greedy iterative and optimal first approaches of feature selection. Out of a total of eight qualities, only four were utilised. Age, BMI, diabetic pedigree functioning, and plasma glucose content were the four characteristics that were considered. The Hoeffding tree method was successful in achieving a recall score of 76.2 percent and an accuracy value of 75.7 percent. (Swapna, et al., 2018) used convolution neural networks with long short-term memory on electrocardiograms. The dataset used was private and comprised of 142000 samples. The researchers were able to reach an accuracy of 90.9 percent using their methods. This dataset did not undergo any pre-processing, nor has it been subjected to any kind of feature selection procedure.

As per (Vasapalli, et al., 2021) diabetes mellitus type 2 (DM) is a condition that may endure for a very long time and whose prevalence has been continuously rising all over the globe. Diabetes affects around 30 million people in India, with millions more at risk for developing the condition. Therefore, early identification is essential in order to prevent diabetes and the difficulties that are linked with it. The purpose of utilising multiple techniques for the hypothetical perseverance of type 2 diabetes rooted on the indicative research is to prolong the disease's detection period by evaluating evocative features and regular practises. As a result, this

will enable the assessment of type 2 diabetes without the usages of clinical exams via the usages of predictive modelling (Vasapalli, et al., 2021).

At this point in time, there is an abundance of clinical information accessible about viruses, their indications, the factors that contribute to sickness, and the effects that they have on one's health. The precision of these algorithms allows for the possibility of accurately predicting the risk of developing type 2 diabetes, which is of vital importance to the medical industry.

(Lekha & Suchetha, 2018) developed their own dataset, which was centred on breathing patterns and included a total of 25 patients. Eleven of these patients seemed normal, five were diagnosed with type 1 diabetes, and the other nine were diagnosed with type 2 diabetes. For the purpose of verification, leave one out cross validation was utilised, and the ROC curve served as the assessment metric. The accuracy of the test was close to 96 percent. (Mohebbi, et al., 2017) employed a CNN and a MLP to identify diabetes on a collection that comprised of 9 individuals. The dataset was given to the researchers to analyse. Constant glucose tracking signal dataset was used as the basis for this dataset. There was a total of nine patients used throughout the study: six for training and validating purposes, three for actual testing. The traditional neural network was able to attain the maximum level of accuracy, which was 77.5 percent.

Pima Indians Diabetes Dataset and Luzhou dataset were both gathered from a regional Chinese hospital by (Zou, et al., 2018) who then used two different feature selection techniques on each dataset. On both sets of data, three different machine learning classifiers—namely, random forest, and neural network—were put through their paces. PCA and minimal redundancy maximal relevance are the names of the feature selection approaches that were used in order to cut down on the total amount of characteristics. Utilizing random forest and the minimal redundancy maximum relevance technique, we were able to reach the highest level of accuracy possible, which was 77.21 percent.

A. Summary

This section summarised all the papers that have predicted diabetes and a small summary of the discussion is presented below in table 1.

S. no.	Method name	Number of datasets used	Name of the dataset	Data size	Speed	Does it rank features	CV protocol used	Evaluation parameters taken	Classifier used	Feature selection method	Number of features used	Classification accuracy	Year in which paper was published	Temporal interval
1	Kamrul Hasan	1	PIDD	768	Slow	Yes	5	Sn, Sp, and AUC	KNN, DT, RF, MLP, AB, XB, and NB	PCA, ICA, and CRB	6	78.9%	April 2020	2020-2021
2	Quan Zou	2	Luzhou and PIDD	68994, 768	Slow	Yes	5	Sn, Sp, ACC, and MCC	J48, RF, and NN	PCA and mrMR	11, 7	80.84%	November 2018	2018-2019
3	Nishith Kumar	1	PIDD	768	Fast	No	5 and 10	Sn, Sp, ACC, PPV, and NPV	GPC, LDA, QDA, and NB	Kernels	All	81.97%	December 2017	2016-2017
4	Maniruzzuman	1	NHANES	9858	Slow	No	2, 5, and 10	Sn, ACC, PPV, NPV, FM, and AUC	NB, DT, RF, and AB	LR	All	92.75%	January 2020	2020-2021
5	V. Jackins	1	PIDD	768	Fast	Yes	None	ACC	NB and RF	CRB	4	74.46%	November 2020	2020-2021
6	N. Sneha	1	PIDD	2500	Slow	Yes	None	Sn, Sp, ACC, PPV, NPV, PLR, NLR, and DP	SVM, RF, NB, DT, and KNN	CRB	11	82.3%	February 2019	2018-2019
7	S. Mohapatra	1	PIDD	768	Fast	No	None	ACC, TP, and TN	MLP	None	All	77.5%	September 2019	2018-2019
8	D. Sisodia	1	PIDD	768	Fast	No	None	Recall, precision, and ACC	NB, SVM, and DT	None	All	76.3%	December 2018	2018-2019

Table 1: Comparative analysis of some papers

III. DISCUSSION AND CHALLENGES

While there have been several approaches to predicting diabetes that rely on a few machine learning techniques like random forests and support vector machines (SVMs), only a few characteristics are picked for prediction. While reading through all of these papers, the researcher encountered the following difficulties:

1) Because the publicly accessible information comprises just nine features, one of which is the class attribute, there was a substantial hurdle in the prediction purpose. Effort and resources are being invested on characteristics that have little predictive value.
2) As the size of the dataset reduces, several authors have removed incomplete data from the real dataset, which might have an impact on the findings.

3) Deep learning and the recurrent neural network have not been used by the authors. As a consequence, a new system must be devised that is more accurate, faster, and more successful in predicting the future.

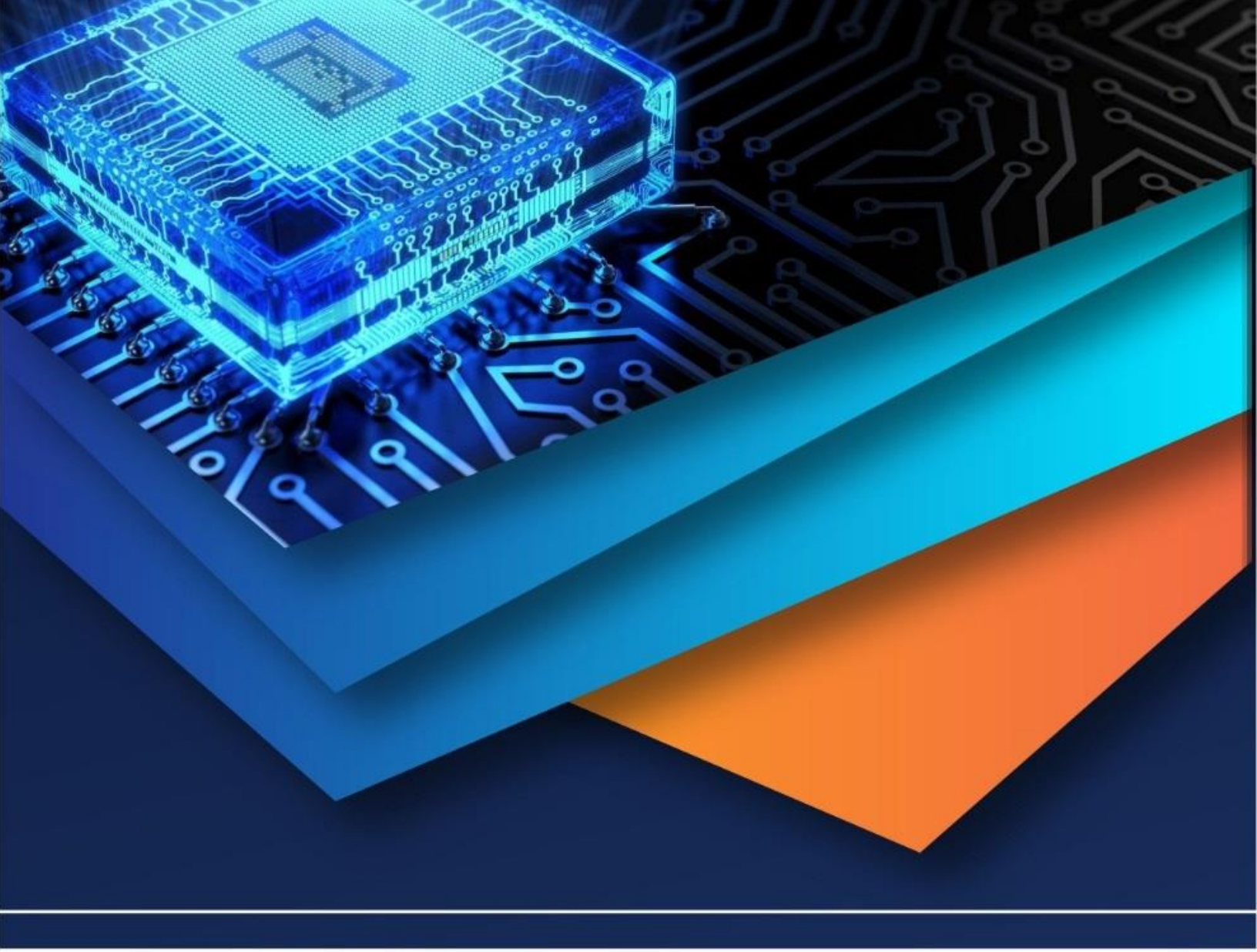
IV. CONCLUSION

In conclusion, the best approach for diabetes prediction was done by (Choubey, et al., 2020) as they used two feature selection approaches called PCA and linear discriminant evaluation to extract relevant characteristics from the PID Dataset. In addition to that, a comparative study of the strategy for selecting attributes was included in the paper. For the aim of classification, a select group of machine learning methods, including radially foundation kernel, KNN, and AdaBoost, were also used to the dataset in question.

REFERENCES

[1] Choubey, D. K. et al., 2020. Comparative analysis of classification methods with PCA and LDA for diabetes. Current Diabetes Reviews, 16(8), p. 833–850.
[2] Gujral, S., 2017. Early diabetes detection using machine learning: a review. International Journal for Innovative Research in Science & Technology, 3(10), pp. 45-60.
[3] Gupta, S., Verma, H. K. & D. Bhardwaj, 2021. Classification of diabetes using naïve bayes and support vector machine as a technique. Singapore, Springer, p. 365–376.
[4] Han, J., Rodriguez, J. C. & M. Behesti, 2020. Discovering Decision Tree-Based Diabetes Prediction Model. Jeju Island, Korea, Springer, p. 99–109.
[5] Hussain, A. & Naaz, S., 2021. Prediction of diabetes mellitus: comparative study of various machine learning models. Advances in Intelligent Systems and Computing, Volume 1166, p. 103–115.
[6] Kandhasamy, J. P. & Balamurali, S., 2015. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, Volume 47, p. 45–51.
[7] Karan, O., Bayraktar, C., Karlık, H. & Karlik, B., 2012. Diagnosing diabetes using neural networks on small mobile devices. Expert Systems with Applications, 39(1), p. 54–60.
[8] Kumari, S., Kumar, D. & M. Mittal, 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering, p. 40–46.
[9] Lekha, S. & Suchetha, M., 2018. Real-time non-invasive detection and classification of diabetes using modified convolution neural Network. IEEE Journal of Biomedical Health Information, Volume 22, p. 1630–1636.
[10] Malik, S., Harous, S. & El-Sayed, H., 2020. Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. Algeria, Springer, p. 95–106.
[11] Mamuda, M. & Sathasivam, S., 2017. Predicting the survival of diabetes using neural network. Poland, AIP Conference Proceedings, p. 40–46.
[12] Mercaldo, F., Nardone, V. & Santone, A., 2017. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science, Volume 112, p. 2519–2528.
[13] Mohebbi, A. et al., 2017. A Deep Learning Approach to Adherence Detection for Type 2 Diabetics. Korea, s.n., p. 2896–2899.
[14] Negi, A. & Jaiswal, V., 2016. A First Attempt to Develop a Diabetes Prediction Method Based on Different Global Datasets. Waknaghat, India, s.n., p. 237–241.
[15] Olaniyi, E. O. & Adnan, K., 2014. Onset diabetes diagnosis using artificial neural network. International Journal of Scientific Engineering and Research, Volume 5, p. 754–759.
[16] Perveen, S., Shahbaz, M., Guergachi, A. & Keshavjee, K., 2016. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, Volume 82, p. 115–121.
[17] Qawqzeh, Y. K. et al., 2020. Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling. BioMed Research International, pp. 1-20.
[18] Rashid, T. A., Abdulla, S. M. & Abdulla, R. M., 2016. Decision support system for diabetes mellitus through machine learning techniques. International Journal of Advanced Computer Science and Applications, 7(7), pp. 1-30.
[19] Saxena, R., Sharma, S. K., Gupta, M. & Sampada, G. C., 2022. A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey. Journal of Healthcare Engineering, pp. 1-22.
[20] Sisodia, D. & Sisodia, D. S., 2018. Prediction of diabetes using classification algorithms. Procedia Computer Science, Volume 132, p. 1578–1585.
[21] Soltani, Z. & Jafarian, A., 2016. A new artificial neural networks approach for diagnosing diabetes disease type II. International Journal of Advanced Computer Science and Applications, Volume 7, p. 89–94.
[22] Sun, Y. L. & Zhang, D. L., 2019. Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey. Technical Gazette, Volume 26, p. 872–880.
[23] Swapna, G., Soman, K. P. & Vinayakumar, R., 2018. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Computer Science, Volume 132, p. 1253–1262.
[24] Tafa, Z., Pervetica, N. & Karahoda, B., 2015. An Intelligent System for Diabetes Prediction. Budva, Montenegro, s.n., p. 378–382.
[25] Tigga, N. P. & Garg, S., 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, Volume 167, p. 706–716.
[26] Vasapalli, M. et al., 2021. Prediction of Type 2 Diabetes Using Machine Learning algorithms. Pichanur, India, s.n.
[27] Yuvaraj, N. & SriPreethaa, K. R., 2017. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing, 22(1), pp. 1-9.
[28] Zou, Q. et al., 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, Volume 9, p. 515–522.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY