Prediction of Diabetes Using Hybridization based Machine learning algorithm

Rishika Lakshmi Naidu Boddeda
Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh, India-530045.
samiripa@gitam.edu

Ritika Prasad
Maharajah's Institute of Medical Sciences, Nellimarla,
Vizianagaram, Andhra Pradesh, India-535217.
prasadritika7@gmail.com

Shanmuk Srinivas Amiripalli
Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh, India-530045.
shanmuk39@gmail.com

Mukkamala SNV Jitendra
Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh, India-530045.
jitendra.mukkamala@gmail.com

2023 International Conference on Smart Computing and Application (ICSCA) | 979-8-3503-4705-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCA57840.2023.10087491

***Abstract*— Diabetes is often referred to as a metabolic disease. It is a category of metabolic disorders caused by a prolonged elevated amount of sugar in the blood. If diabetes can be predicted early, the chance and danger of diabetes can be minimized. Owing to the small amount of labeled data and the inclusion of outliers (or incomplete values) in diabetes databases, predicting diabetes correctly and accurately is extremely difficult. We propose a rigorous architecture for diabetes prediction that includes outlier exclusion, missing meaning filling, data standardization, function collection, and various Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, Logistic regression, support vector machine). In addition, we suggested the creation of a hybrid algorithm. We used a variety of algorithms to conduct hybridization in order to improve accuracy. Pima Indian Diabetes Data Collection was used as the data source.**

Keywords*— Machine learning, Diabetes, SVM, K-Nearest Neighbour, Decision Tree, Random forest, Logistic regression, PIMA data set.

I. INTRODUCTION

Diabetes is one of the most deadly illnesses on the planet. Diabetes is not just a single illness, but also a trigger for a variety of other conditions such as heart failure, paralysis, renal disease, and so on. Patients often attend a diagnostic centre, meet with their specialist, and wait a day or so for their results. We will address these issues by using machine learning. The key goal of this research is to develop a model that can accurately analyze or forecast the risk of diabetes in patients.

We present diabetes prediction using machine learning classification algorithms in this article, and we will use hybridization to improve accuracy. There are three varieties of diabetes: type 1 diabetes, type 2 diabetes, and type 3 diabetes. Type 1: It can affect someone at any age, although it is more common in adolescents and between the ages of adolescence and adulthood. Type 2 diabetes is more prevalent in teenagers, accounting for about 90% of all diabetes events. When you have type 2 diabetes, the body cannot properly use the insulin that it produces. Tier 3 diabetes is characterized by elevated blood glucose levels during pregnancy and is linked to problems for both the mother and the infant.

Abnormal thirst and dry mouth, sudden weight loss, frequent urination, lack of stamina, tiredness, constant nausea, blurred vision, bedwetting, and sores that do not cure are some of the signs of diabetes. Overweight, an unstable diet, obesity, and other factors contribute to diabetes. One of the most important areas for such forecasts and study is machine learning. According to statistics from the World Health Organization, India has the most diabetes incidents. Data mining is essential for collecting or compiling various types of data and filing it in a manner that it can be conveniently obtained. We train the algorithms with data sets that are used for data testing and prediction utilizing supervised machine learning.

II. LITERATURE SURVEY

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

Different Machine Learning Approaches for Diabetes Prediction Priyanka Sonar and JayaMalini (2019) used a variety of machine learning algorithms, including decision trees, random forests, and artificial neural networks, to improve prediction accuracy[1]. The PIMA dataset was included. To begin, they discovered that they needed a large number of decision trees in order to achieve better precision, and they trained RF and ANN to identify hidden trends in the results, achieving 0.75 accuracy[2]. They said that even unstructured data can be useful for prediction because it allows us to learn new things from established data.T. Abbas and Marely Rios (2019) used a Pima Indian data collection to forecast diabetes in healthy people using machine learning. They used 10-fold cross validation and a help vector machine for prediction, and they used Matlab for machine learning routines and data analysis. They got an accuracy of 0.72 [3].

The forecast models may be used on other related datasets that provide OGTT measurements as part of a future expansion of this research.Md. Maniruzzaman, Md. Jahanur

Rahman, and BenojirAhamad (2018) used classification algorithms for diabetes prediction in their paper Accurate Diabetes Risk Stratification Using Machine Learning [4]. These researchers obtained 0.79 precision utilising machine learning classification models such as naive Bayesian, logistic regression, Decision tree, and support vector machine in 2018. They also reported that pre-processing techniques can be used to substitute irrelevant values with mean or median and outliers with mean or median [5].

There are many other methods for feature extraction, feature filtering, and classification, and the performances of the provided device combinations can be contrasted to those of competing systems. In 2020, conducted research using ensemble methods such as Ada boost and random forest multi layer perceptron, which are mostly used to locate hidden layers and association of the vector and are also used in supervised machine learning, and for better accuracy with different models leaving a domain, while also noting that different combinations of algorithms may be attempted [6] .

Diabetes Prognosis Md. Kamrul Hasan and Md. AshrafulAlam (2020) used both classification and ensemble models in their paper Implementing a Mix of ML Classifiers into an Ensemble. Improved attribute-to-outcome correlation is a key benefit of correlation-based attribute selection, according to Ada boost, XG boost, and multilayer perceptron, which is also used to find the secret layers and correlation of the indicator in controlled machine learning for diabetes prediction. They have used k-fold cross validation and has been reported to increase the link between selected attributes and desired outcomes [7].

Ritu Chauhan and Ashish Kumar Mourya used Gradient Boosting, Logistic Regression, and Naive Bayes in their machine learning method to predict and diagnose potential diabetes risk. They obtained a certain level of precision, and they noted that we can run the same algorithms on various data sets and compare the results, as well as that trying on broad data sets can help us achieve better results.Any machine learning classification methods, such as logistic regression, help vector machine, decision tree, naive Bayesian, random forest, and k-nearest neighbor, have been applied after a review of academic articles. In addition, we suggested a hybrid algorithm for conducting a thorough investigation [8].

Our key goal is to examine a diabetes dataset to determine whether or not medical examination outcomes may be used to forecast diabetes. To look at several current models to see what benefits they provide for diabetes identification. To create a new model based on machine learning and compare it to an established model [9].

In this research, data collected from Kaggle, which is the PIMA Indian dataset. This data collection has been used in a variety of studies, not just for diabetes prediction but also for kidney disease prediction. This data collection includes 768 PIMA Indian dataset patient records with 9 attributes. The data set's description is provided in tabular form below. We begin by pre-processing the data by eliminating unnecessary data and outlier rejection, as well as removing all noise. For our diabetes forecast, we used a variety of classification algorithms [10].

For the experimental analysis, we used a variety of algorithms. Random Forest, Logistic Regression, SVM, Decision Trees, and Naive Bayesian Nearest Neighbor are some of the algorithms used.

A. Random Forest

One of the supervised machine learning algorithms is random forest. Both classification and regression are done with this system. It is a set of trees, and the average prediction of individual trees is taken during preparation and research. Partition our data set into separate subsets using random tree. Each subset's data is distinct from the others, and the trees are then trained on their subsets defined, with the performance expected by the tree with the most votes. To divide the tree, use a random subset of features. The tree's outputs are all independently stored. The increased number of trees aids us in making accurate predictions. In random woodland, each tree grows to the fullest extent without being pruned at first. The outcome is then predicted using the plurality votes from all of the trees. Node depth, number of trees, number of functions, number of threshold values, and so on is all parameters in decision trees. We must be cautious if there are any imbalanced plants, since strongly imbalanced trees have a detrimental impact on forecasts. Confusion matrix for random forest [11].

Tested positive	24
Tested Negative	76

B. Logistic regression

A classification approach is logistic regression. It has a series of guidelines that assist us in making predictions. This divides our performance based on the chance of the class or several groups, rather than concluding output based on a single class. This shows the interaction between the independent and dependent variables. It classifies data in binary type, that is, just in 0 and 1, indicating whether the patient is positive or negative for diabetes. The main goal of logistic regression is to find the best match such that it can describe the relationship between the objective and predictor variables. The linear regression model supports logistic regression. The sigmoid function is used in the logistic regression model to predict the likelihood between positive and negative classes. Sigmoid function P as in (1).

$$P = 1/1 + e - (a + b\,x) \quad (1)$$

Here P = probability, a and b = parameter of Model. Confusion matrix for logistic regression.

Tested positive	32
Tested Negative	68

C. Support vector Machine

A supervised machine learning algorithm called support vector machine is sometimes used. This is mostly used to convert or separate data depending on specific outputs. Both classification and regression can be done using this method. We use a hyper plane to segment our class and measure the difference between the planes, which is known as margin. If the difference between them is small, there is a risk of misunderstanding, so we can choose classes with a large gap for better results. In general, we may use logistic regression to make predictions, but model fit becomes complicated when the data is high dimensional or there are several missing values. Since SVM is a model-free approach,

no assumptions or distributions for the dependent and independent variables are needed. Each data point is defined as an n-dimensional vector on its own [12]. We construct n-1 dimensional hyper planes to divide two groups of the greatest distance and data points on both sides. To transform into a multidimensional domain, we use nonlinear Kernels. In this model, various forms of kernels can be used. Sigmoid kernels, Gaussian kernel, polynomial kernel, and so on are some of them. Confusion matrix for SVM.

Tested positive	24
Tested Negative	76

D. Decision Tree

It's a deep learning algorithm that's supervised. This is described as a tree, with each node corresponding to a different class name. It provides us a statistical likelihood, where each branch of the tree reflects a potential choice, and any internal node of the tree produces a Boolean kind of result. We must prevent over fitting data in decision trees; whether there is some noise or the data collection is minimal, we would have difficulty predicting [13]. Reduced error pruning is used to solve this issue, allowing each node to be substituted by the most critical feature that aids prediction. Continually respected characteristic is what we use. It is an easy and quick procedure. We get a Boolean kind of outcome from any internal node of the tree, which gives us a statistical likelihood where each branch of the tree reflects a potential choice. Confusion matrix for decision tree.

Tested positive	34
Tested Negative	66

E. Naive Bayesian

The Naive Bayesian algorithm is a supervised machine learning method. One of the most widely employed classification algorithms is this one. This allows for forecasts based on various attributes. It makes predictions about features that aren't related to one another. It is a widely used algorithm that predicts the class of a test dataset in a short amount of time [14]. The advantage of naive Bayesian is that it is plain and straightforward to use. The disadvantage is that the outcome is more accurate owing to a higher likelihood value. Strong expectation about the data distribution's form. There has been deterioration in precision. The naive Bayesian algorithm employs a smoothing method. It's a data-cleaning procedure that eliminates zero values. It is one of the best algorithms for classification problems since it is dependent on conditional probabilities. Where the data is strongly imbalanced, naive Bayesian may be used to solve the dilemma. And if the data contains noise or incomplete values, this algorithm can be useful in coping with these situations [15]. Confusion matrix for naive Bayesian.

Tested positive	28
Tested Negative	72

F. K-Nearest Neighbour

KNN is a deep learning algorithm that is supervised. It's an algorithm that uses a series of rules to store all of the data in a data collection. The remainder of the data is separated into groups, and we need to locate the closest neighbor to our classes. We use the balance of the label's votes to determine the winner.K is the number of adjacent neighbors, and it is often a positive integer in this case. The meaning of a neighbor is selected from a list of classes. The Euclidean distance is used to describe proximity. The equation defines the Euclidean distance between two points P and Q, i.e. P (p1,p2,..Pn) and Q (q1,q2,..qn).KNN keeps track of all instances and lets us identify them based on how close the characteristics are. How does it distinguish between two things? KNN looks for nearby points that share similar characteristics and divides them into two groups based on those characteristics. It also creates additional divisions based on the characteristics of the data collected. It is a parameter that corresponds to the number of closest neighbors that the majority voting method requires. It classifies new data points based on their proximity to their closest neighbours. Confusion matrix for KNN.

Tested positive	26
Tested Negative	73

III. DATA PRE-PROCESSING

Data mining is typically used to store or gather information. Most health-care-related data collected contains missing values, unnecessary data, and other impurities, so data pre-processing is performed to increase data accuracy. This method is critical in ensuring that machine learning models perform well for our data collection. We're working with the PIMA Indian dataset from Kaggle. There are two stages to this pre-processing: Cleaning Up: Getting Rid of Null Values (0). A value of zero cannot be obtained. This occurrence is therefore now false. When a data set has the most number of missing values, the whole row is removed when an accumulation of missing values contributes to misclassification. We render function subsets by removing unnecessary functions/instances, a method known as features subset filtering, which decreases the dimensionality of data and allows us to operate faster. Both outliers and unnecessary data or noise are omitted from your key data while you use this function selection. The data may not be completely excluded from the framework because the unwanted data omitted from it might be valuable for other forecasts, so the data will simply be discarded and a subset will be created. Data normalisation occurs during model training and validation, following data splitting. After the data has been partitioned, it is used to train an algorithm while the test data is ignored. Both grouping and regression questions use the train and test sets [16].

IV. PROPOSED WORK

The hybridization algorithm is proposed in this study. In order to improve precision, we conducted hybridization on various combinations of algorithms. Combining or combining two separate algorithms is referred to as a hybrid algorithm. To improve efficiency, the features of two algorithms are

merged. In this research, Kaggle PIMA Indian Diabetes data package was used. We conducted various operations on our data set by using various parameters. To obtain the accuracy performance, we educated our dataset using various algorithms.This is the most crucial step, which involves the creation of a diabetes prediction model. We used different machine learning algorithms for diabetes prediction, which were addressed earlier. In our proposed methodology many combinations were tried but some of them were mentioned in "Fig.1":

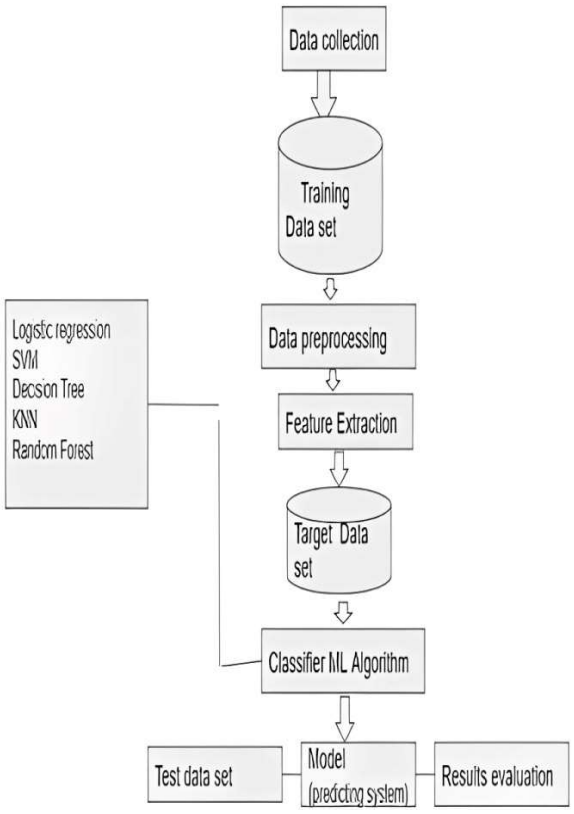

Fig. 1. Proposed Hybridization model

HYBRIDIZATION ALGORITHM

Step 1: Begin

Step 2: Import the required libraries and the diabetes dataset.

Step 3: Delete any missing info.

Step 4: The data collection is divided into two parts: training and testing.

Step5: Choose a hybrid algorithm, which is a combination of two machine learning algorithms (K- Nearest Neighbour, Support Vector Machine, Decision Tree, Logistic regression, Random Forest).

Step6: Using the training set, a classifier model is created for the described machine learning algorithm.

Step 7: Using the test collection, test the Classifier model for the described machine learning algorithm.

Step 8: Make a comparison The experimental output outcomes for each classifier are evaluated.

Step 9: Determine the highest performing algorithm after testing it using different metrics.

Step10.Stop

V. RESULT ANALYSIS

Along with the proposed work, a comparison of six related algorithms is performed. The uncertainty matrix is used to measure the algorithm's accuracy.

The subset must first match the model, after which predictions are produced and compared to the current and proposed algorithms based on their results. The software is fitted with the train data collection, and the model is evaluated with the test data set. Traditional scalar method-The Standard The data is transformed by the scaler process into a normal distribution with zero mean and one standard deviation. When dealing with multivariate outcomes, this is accomplished in a feature-by-feature fashion. (independently for each column of the data). When we get more harmful qualities, this is more likely to be used. X is equal to x-u/sigma [17].

The mean of the function values is Mu, and Scaling of features: The standard deviation of the function values is referred to as Sigma. The values in this case are not restricted to a specific number. Since the regular scalar performs the same function as the normalize (transforming data from 0 to 1), it is more efficient in classification than regression and vice versa. Accuracy is measured by the formula as in (2).

$$Accuracy = \frac{TP+T}{TP+TN+FP+F} \quad (2)$$

TABLE I. COMBINATION OF MODELS AND ITS ACCURACIES.

Algorithm	*Accuracy*
Logistic regression and SVM	0.79
Naive Bayesian and SVM	0.80
KNN and Logistic regression	0.81
Random forest and KNN	0.80
KNN and SVM	0.85

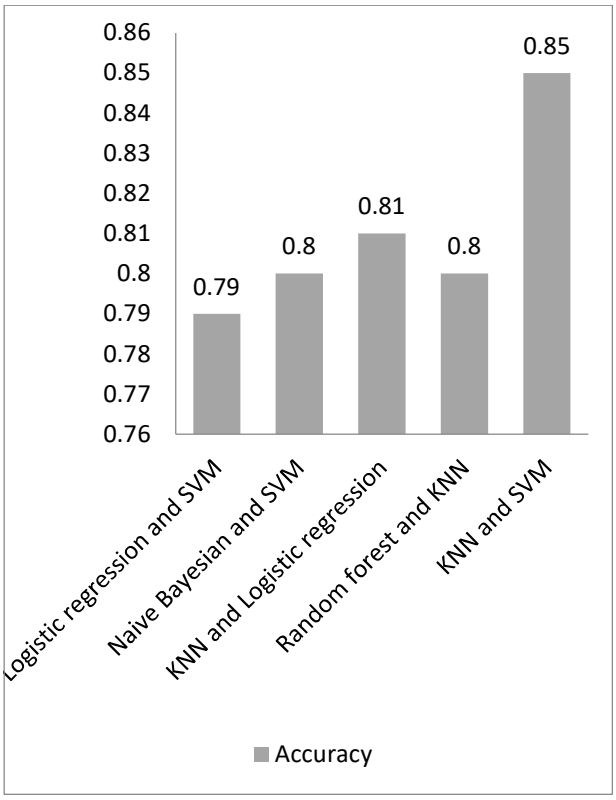

Fig. 2. Comparison of hybrid algorithms.

TABLE II. COMPARISION OF EXISTING MODELS WITH PROPOSED HYBRIDIZATION ALGORITHM.

Algorithms	*Accuracy*
Logistic Regression	0.77
Support Vector Machine	0.78
Decision Tree	0.72
Random Forest	0.74
K-Nearest Neighbor	0.77
Proposed work (Hybrid algorithm)	0.85

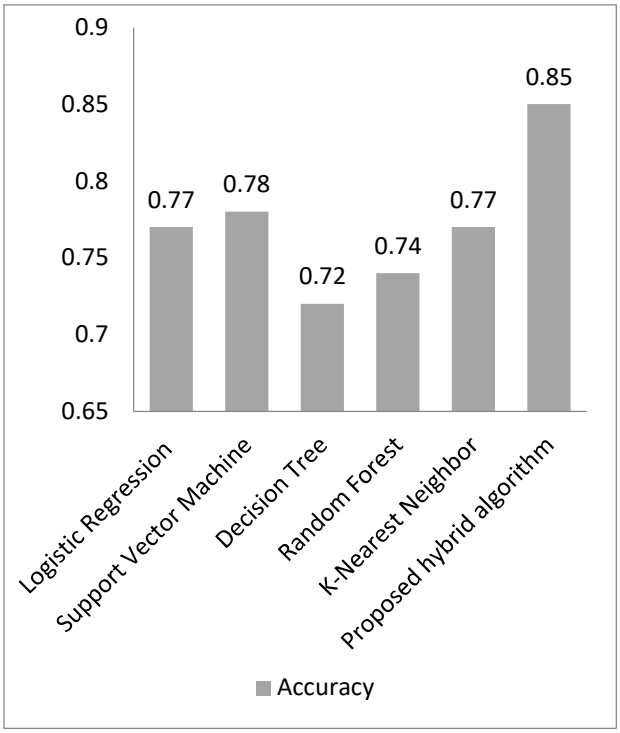

Fig. 3. Comparison of hybrid algorithms

VI. CONCLUSION

One of the most difficult things is diabetes prediction. We first tested the current approach to see how well it performed, and then we combined two algorithms to improve accuracy. We trained the data on various models using the PIMA Indian dataset after pre-processing it to see how it performed. We used six separate algorithms, with the best yielding an accuracy of 0.78. We suggested a hybrid algorithm to improve precision by combining various algorithms. We observed that using a KNN and SVM hybrid algorithm yielded a result of 0.85.

ACKNOWLEDGMENT

We acknowledge support from the GITAM (Deemed to be University), Vishakhapatnam and Maharajah's Institute of Medical Sciences, Vizianagaram, for continuous support, valuable suggestions, and very useful discussions for all the support being extended to carry out this research work.

REFERENCES

[1] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of medical systems, 42(5), 1-17.

[2] Syed, A. H., & Khan, T. (2020). Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. IEEE Access, 8, 199539-199561.

[3] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[4] Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8(1), 1-14.

[5] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning Algorithms in healthcare. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

[6] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.

[7] Abbas, H., Alic, L., Rios, M., Abdul-Ghani, M., & Qaraqe, K. (2019, June). Predicting diabetes in healthy population through machine learning. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 567-570). IEEE.

[8] Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1(9), 1-8.

[9] Maniruzzaman, Md, et al. "Classification and prediction of diabetes disease using machine learning paradigm." Health information science and systems 8.1 (2020): 1-14.

[10] Reddy, G. T., Bhattacharya, S., Ramakrishnan, S. S., Chowdhary, C. L., Hakak, S., Kaluri, R., & Reddy, M. P. K. (2020, February). An ensemble based machine learning model for diabetic retinopathy classification. In 2020 international conference on emerging trends in information technology and engineering (ic-ETITE) (pp. 1-6). IEEE.

[11] Verma, A. K., Pal, S., & Kumar, S. (2020). Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study. Applied biochemistry and biotechnology, 190(2), 341-359.

[12] Xie, J., & Wang, Q. (2020). Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison With Classical Time-Series Models. IEEE Transactions on Biomedical Engineering, 67(11), 3101-3124.

[13] Islam, M. T., Raihan, M., Aktar, N., Alam, M. S., Ema, R. R., & Islam, T. (2020, July). Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[14] Amiripalli, S. S., Kollu, V. V. R., Jaidhan, B. J., Srinivasa Chakravarthi, L., & Raju, V. A. (2020). Performance improvement model for airlines connectivity system using network science. International Journal of Advanced Trends in Computer Science and Engineering, 9(1), 789-792.

[15] Potharaju, S. P., Sreedevi, M., & Amiripalli, S. S. (2019). An Ensemble Feature Selection Framework of Sonar Targets Using Symmetrical Uncertainty and Multi-Layer Perceptron (SU-MLP). In Cognitive Informatics and Soft Computing (pp. 247-256). Springer, Singapore.

[16] Thota, J. R., Kothuru, M., & Shanmuk Srinivas, A. Monitoring Diabetes Occurrence Probability Using Classification Technique With A UI.

[17] Jaidhan, B. J., Madhuri, B. D., Pushpa, K., & Devi, B. L. Application of Big Data Analytics and Pattern Recognition Aggregated With Random Forest for Detecting Fraudulent Credit Card Transactions (CCFD-BPRRF).