EasyChair Preprint

№ 10686

Diabetes Prediction Using Machine Learning Approach

V Viswanatha, A.C Ramachandra, Dhanush Murthy and
H Thanishka

August 7, 2023

Diabetes Prediction Using Machine Learning Approach

Viswanatha V[1], Ramachandra A.C[2], Dhanush Murthy [3] and Thanishka [4]
[1]Asst.Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India
[2]Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India
[3,4]Student, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India

[1]Corresponding Author: Viswanatha V

Abstract -Diabetes is one of the most common diseases in the world, when detected early, it is possible to stop the progression of the disease and prevent further complications. In this work, we design a predictive model that predicts whether a patient will develop diabetes, based on certain diagnostic measures contained in the dataset, and explore different techniques to improve performance and accuracy. Logistic regression is the main algorithm used in this article and the analysis was performed using Python IDEs. The trial mainly uses two data sets one is the PIMA Indians Diabetes dataset, which is the source from the National Institute of Diabetes and Digestive and Kidney Diseases, and another dataset from Vanderbilt, based on a study of rural African Americans in Virginia. The selection of functions is done using two different methods. Aggregation methods are used in addition, which improves performance by producing better prediction against a single model. Accuracy and runtime are recorded for the original datasets and for those obtained later using feature selection and aggregation techniques. A comparison is also presented in each case. Highest accuracy obtained is about 78% for dataset 1, after using aggregation technique - Maximum Voting; and it was around 93% for dataset 2, after using combined techniques: maximum polling and stacking. Logistic regression has been proven to be one of the effective algorithms for building predictive models. This study also shows that in addition to algorithm selection, there are other factors that can be improved model accuracy and runtime, such as: preprocessing data, removing redundant and null values, normalization, cross-validation, feature selection, and the use of aggregation techniques.

I. INTRODUCTION

Diabetes, also known as diabetes mellitus, affects many people around the world According to the International Diabetes Federation About 463 million adults (aged 20-79) had diabetes in 2019. They predicted that number will increase to 700 million by 2045. The prevalence of diabetes has increased more rapidly in low- and middle-income countries than in high-income countries. Diabetes is the main one cause of blindness, kidney failure, heart attack, stroke and lower limbs amputation It is also believed that about 84.1 million Americans People over the age of 18 have prediabetes There are three types of diabetes. Type 1 is known as insulin-dependent diabetes mellitus (IDDM). The reason for this type Diabetes is the inability of the human body to produce enough insulin. in In this case, the patient must inject insulin. Type 2 is also known as non-insulin dependent diabetes mellitus (NIDDM). Anyway, Diabetes occurs when the body's cells are unable to use insulin properly.

Type 3 gestational diabetes increases blood sugar in pregnant women. This happens when diabetes is not detected at an early stage. Although diabetes is incurable, it can be treated with treatment and medication. Many healthcare organizations are now using machine learning techniques in healthcare, such as predictive modelling. In addition, the game has sophisticated algorithms that recognize processes and patterns invisible to the human eye. This helps researchers find new drugs and treatment plans. Predictive modelling uses data mining, machine learning and statistics to identify patterns in data and identify opportunities to realize results. This article focuses on building a diabetes prediction model to determine whether a particular patient has diabetes, and then explores various techniques to improve accuracy.

II. LITERATURE REVIEW

Machine learning techniques are increasingly useful in the field of medicine. Many researchers have used various machine learning and deep learning techniques and algorithms to predict diabetes. Aishwarya and Vaidehi used several machine learning algorithms like support vector machines, random forest classifier, decision tree classifier, extra tree classifier, Ada boost algorithm, perceptron, linear discriminant analysis algorithm, logistic regression, K-NN, Gaussian naive gulf, Baging Algorithm. Bagging and gradient boosting classifier. To test the different models, they used two different datasets - PIMA India and another Diabetes dataset. Logistic regression gave them 96% accuracy. On the other hand, Tejas and Pramila chose two algorithms - Logistic Regression and SVM - to build a predictive model for diabetes. Data processing was done to get better results. They found that SVM performed better with 79 percent accuracy. Three different machine learning algorithms – Random Forest, Decision Tree and Naive Bayes are used to built ML model. Material pre-processing techniques are used. The results showed that the highest accuracy of 94% was obtained by the Random Forest algorithm. Deepti and Dilip used decision tree, SVM and Naive Bayes algorithms. Tenfold cross-validation was used to improve performance. The highest accuracy was achieved by Naive Bayes, with 76.30 percent accuracy. Both articles used the Pima Indian Diabetes database. Deep learning methods for diabetes prediction. The first used a multilayer Feed-Forward neural network. A backpropagation algorithm was used to train the model. They also used the PIMA-India dataset and normalized it before preprocessing to obtain numerical stability. Their accuracy was 82. The latter used the dataset. The dataset consisted of 142,000 samples and eight attributes. They achieved 93.6 percent accuracy with the CNN model and 95.1 percent accuracy with the CNN-LSTM model, and five-fold cross-validation for both. All the above studies provided a comparative performance analysis of different machine learning algorithms. Some of them used data preprocessing and cross-validation techniques to improve accuracy, but they all focused more on comparing performance different models instead of improving one model. In this article, I focused on one model and explored techniques that not only improve accuracy, but also improve execution speed, thus increasing performance. This article shows that, in addition to the choice of algorithms, data pre- and post-processing play an important role. general improvement of the model.

III. MODEL DESIGN

This article uses two main datasets. The first is the PIMA India dataset, which consists of 768 patients, all of whom are women at least 21 years of age. There are nine functions in total. Another dataset comes from Vanderbilt, based on research in rural Africa Americans in Virginia. It consists of 16 functions. There are 390 data samples from both male and female patients. Here dataset 1 refers to the PIMA India dataset and dataset 2 refers to the Vanderbilt dataset. The main algorithm used in the predictive model is logistic regression, although only a few other machine learning techniques such as decision trees, support vector machines, K-nearest neighbours and Naive Bayes are used in ensemble methods to test the improvement of the original performance. We constructed a flowchart of the predictive model shown in, It seems to apply current is implemented. Various methods are being explored to improve performance and execution time. First, it starts with two methods of feature selection - creating new features and then selecting the best ones; and another method is one-dimensional feature selection. Another technique is to use ensemble methods. This project uses two ensemble methods, namely maximum voting/majority voting and stacking. All results are analysed using IDE PyCharm and Python 3.6 on a Windows 10 platform.

IV. METHODS

The first steps are the selection of the model dataset and the evaluation of its characteristics. The first dataset selected for this paper is the PIMA India dataset. There are a total of nine characteristics/variables, eight of which are predictor variables and one target variable. Features are as follows:

- Pregnancy: several times the patient has been pregnant.
- Glucose: plasma glucose concentration over two hours orally glucose tolerance test.
- Blood pressure: diastolic blood pressure (mm Hg).
- Skin thickness: Triceps skinfold thickness (mm).
- Insulin: two-hour serum insulin (mu U/ml).
- BMI: body mass index (weight in kg/(height in meters)^2).
- Diabetes Pedigree Function/DPF: A function that calculates the probability diabetes based on family history.
- Age: in years.
- Outcome: Categorical variable (0 if not diabetic, 1 if diabetic). it is target variable.
- Another dataset used is the Vanderbilt dataset. It consists of 16 characteristics, one of which is the target variable, i.e., diabetes:
- Patient Number: Identifies patients by number
- Cholesterol: total cholesterol
- Glucose: fasting blood sugar
- HDL: HDL or good cholesterol
- Age: Age of the patient
- Gender: 162 men, 228 women
- Height: inches

- Weight: lbs (lbs)
- BMI: 703 x weight (lbs)/ [height (in)2
- Systolic blood pressure: the top number of the blood pressure
- Diastolic blood pressure: lower blood pressure
- Waist: Measured in inches
- Hip: Measured in inches
- Waist/hip: ratio is a potentially stronger risk factor for heart disease than BMI
- Diabetes: yes (60), no (330)

Data in CSV format is loaded into a variable. Dataset 1 has 768 data points and Dataset 2 has 390 data points.
Data mining involves extracting data and finding relationships between features. Data set 1 consists of 268 diabetic patients and the remaining 500 non-diabetic patients, while data set 2 contains 60 diabetic and 330 non-diabetic patients. The heatmap shown on the screen shows the relationship between the features of dataset 1. Lighter colors represent higher correlation and darker colors represent lower correlation. IT shows a bar chart shows the number of patients with or without diabetes in dataset 1.

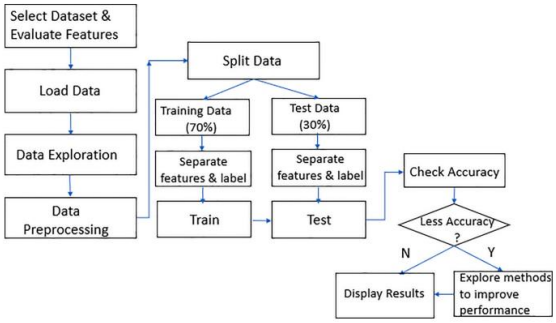

Fig. 1. Flowchart of the Diabetes Prediction Model

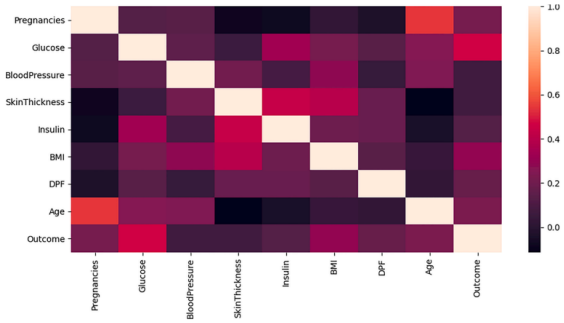

Fig. 2. Correlation values and heat map, showing the correlation between the features for Dataset 1

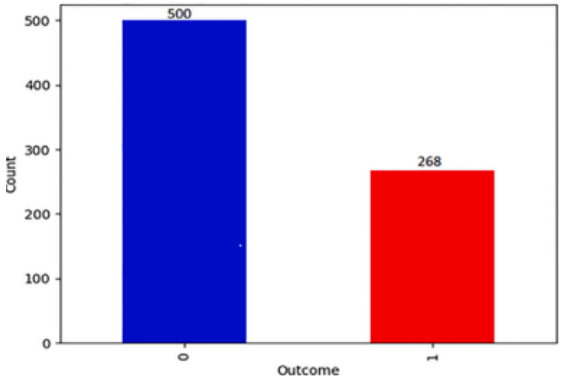

Fig. 3. Count of patients with and without diabetes

A).DATA PREPROCESSING

Missing values, null values, and non-null values for predictor variables must be identified in the dataset. Predictor variables/functions cannot have a zero value, except for certain functions, such as the Pregnancy function in dataset 1. These values must be replaced by the average values of the column. This is an important step to increase the accuracy of the prediction, because incorrect values increase the probability of incorrect predictions.

B).TRAINING AND TESING

The data is divided into training and test sets. Common ratios are 80:20 and 70:30. In this project, the data is divided in a ratio of 70:30, ie. 70% for training and 30% for testing. A logistic regression algorithm is used to make predictions and check accuracy. Complete time is also calculated. There are four important terms for predictions: - True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP and TN represent the cases where the actual result and the result are the same, while FP and FN are the cases where the opposite results are obtained. A classification report is produced that includes precision, recall, F1 score and support. The accuracy meter shows how many percent predictions are correct. Recall describes what percentage of positives are correctly identified. The F1 score is the percentage of correct positive predictions. The support is the number of actual instances of the class specified data set.

Table 1 shows the classification report for both the
datasets:

Precision = TP/(TP + FP)

Recall = TP/(TP+FN)

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Table 1
Classification report.

For Dataset 1	Precision	Recall	F1 score	Support
0	0.77	0.85	0.81	151
1	0.66	0.53	0.58	80
Accuracy			0.74	231
Macro Avg.	0.71	0.69	0.70	231
Weighted Avg.	0.73	0.74	0.73	231
For Dataset 2	Precision	Recall	F1 score	Support
0	0.90	0.96	0.93	93
1	0.78	0.58	0.67	24
Accuracy			0.88	117
Macro Avg.	0.84	0.77	0.80	117
Weighted Avg.	0.87	0.88	0.87	117

V. RESULTS AND DISCUSSIONS

Various methods are used to improve accuracy. The methods include: - creating new functions; feature selection using one-dimensional feature selection; and ensemble methods such as Max Voting and Stacking.

Five new features are created from the existing features and added to the dataset. These features were created based on research into specific diagnostic measurements for diabetes. A heatmap is created and eight characteristics are selected for the output and each other based on the correlation. Thanks to these eight qualities accuracy is recalculated. Accuracy seemed to increase. This method uses dataset 1. The new features are labeled NF1, NF2, NF3, NF4 and NF5. The first parameter, NF1, is basically chosen on the basis that people over 30 are generally less susceptible to the disease. A blood sugar level below 140 mg/dl is also normal. NF2 is for people with BMImore than 30 kg/m2, who have a higher risk of developing diabetes. NF3 is selected by Chengjie Lv et al. based on research done by women with four or more pregnancies have a higher risk of developing diabetes than women with three or fewer pregnancies. Normal diastolic blood pressure is less than or equal to 80 and it is The most important factor in NF4. The last characteristic of NF5 is the combination of normal glucose and a higher BMI value. These features are described below.

- NF1 - age less than or equal to 30 and glucose level not more than 140
- NF2 - BMI less than or equal to 30
- NF3 - age less than or equal to 30 and less than or equal to three pregnancies
- NF4 - glucose level below or equal to 140 and lower blood pressure than 80 or equal
- NF5 - Glucose value less than or equal to 140 and BMI less than or equal to 45 Correlation between all old and new characteristics is calculated.

It shows a heatmap with characteristic correlation values. Depending on the correlation values, the eight features are chosen to have less correlation with each other and more correlation with the output. Additional features are removed and only the selected eight features are retained. Figure 5 is a heatmap with only stored values. Again, training and testing are done with these new features, and accuracy and execution time are considered. Table 2 shows the classification report with new features. Table 3 also shows the difference in execution times and accuracy before and after using this method. We see an increase in accuracy and a big change in program execution time, which has been significantly reduced.

Table 2
Classification report for dataset 1 after Feature Selection.

	Precision	Recall	F1 score	Support
0	0.81	0.82	0.81	151
1	0.65	0.62	0.64	80
Accuracy			0.75	231
Macro Avg	0.73	0.72	0.73	231
Weighted Avg	0.75	0.75	0.75	231

Table 3
Comparison before and after feature selection for Dataset 1.

	Accuracy	Execution Time
Before	0.7403	0.042
After	0.7532	0.008

Table 4
Classification report for dataset 2 after univariate feature selection.

	Precision	Recall	F1 score	Support
0	0.89	0.98	0.93	93
1	0.87	0.54	0.67	24
Accuracy			0.89	117
Macro Avg	0.88	0.76	0.80	117
Weighted Avg	0.89	0.89	0.88	117

A).UNIVARIATE FEATURE SELECTION

A technique called univariate feature selection chooses the best features based on univariate statistical tests. Every attribute is contrasted with the target variable to identify any statistically significant relationships between them. The term "analysis of variance" is sometimes used. The Univariate feature selection is specifically used to carry out the feature selection, together with the chi-square test. Chi-Square calculates the difference between the predicted and actual counts. The following is the formula:

(c)2 equals (Oi Ei)2 /Ei

where: c = degrees of freedom

O = observed value(s)

E = expected value(s)

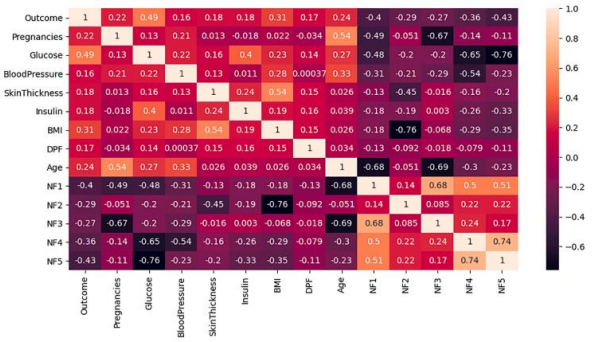

Fig. 4. Heat map showing the correlation between all the old and new features.

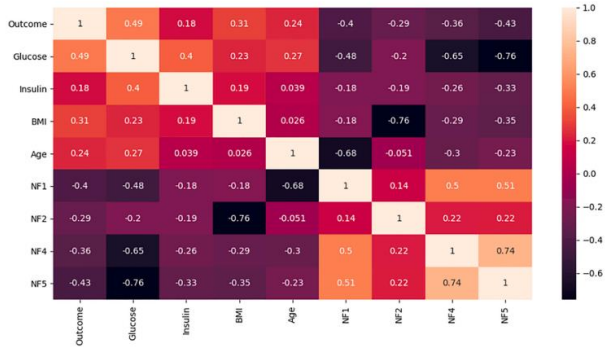

Fig. 5. Heat map with correlation values of retained features.

B).ENSEMBLE METHODS

By mixing many models, ensemble methods/algorithms assist to enhance the outcomes of machine learning. Therefore, by employing such techniques, the performance of the prediction models may be further enhanced. Many renowned machines learning contests, including the Netflix Competition, KDD 2009, and Kaggle, awarded the top prizes to ensemble approaches. In this project, two techniques—Max Voting and Stacking—are utilised to boost performance, and the variations that result from their application are recorded. One of the simplest ensemble procedures is the maximum voting/majority voting. Predictions from several machine learning algorithms are combined. Each model's predictions are referred to as a "vote." The final forecast is derived from the predictions that received most model votes.

P(X) = argmaxi ∑N j=1wj pijwhere: P(X) is the final prediction wj is the weight of the jth classification pij is the is the probability estimate from the jth classification rule for ith class.

Table 5
Comparison before and after feature selection for dataset 2.

	Accuracy	Execution Time
Before	0.8803	0.036
After	0.8889	0.005

Results summary: After pre-processing the null values and removing the missing data, the predictive model was first created using simply the Logistic Regression technique. Later, the accuracy and execution time were improved by using the feature selection approaches. For Dataset 1, existing features were used to construct new ones, and the top eight features were chosen using a correlation heat map. Eight of the highest-scoring features from Dataset 2 were selected using univariate feature selection with the chi-square test. Further attempts were made to improve performance by using ensemble techniques. On both datasets, the Max/Majority Voting and Stacking approaches were evaluated. The previous approach shown a considerable improvement in performance, making it the greatest way of all. Following cross-validation, the latter performed well.

Table 7
Comparison of Accuracy with and without Stacking.

	Base Models	Accuracy after k-fold	Final Model Accuracy of Log Regression after k-fold
Dataset 1	Decision Tree	0.724	
	Naïve Bayes	0. 748	0.7635
	k-NN	0. 756	
Dataset 2	Decision Tree	0.923	
	Naïve Bayes	0.930	0.9304
	k-NN	0.923	

Table 6
Accuracy for individual and ensemble models.

	Logistic Regression	Decision Tree	Support Vector Machine	k-Nearest Neighbor	Naïve Bayes	Ensemble Model
Dataset 1	0.7532	0.7229	0.7489	0.6797	0.7403	0.7783
Dataset 2	0.8889	0.8803	0.8803	0.8803	0.8889	0.9341

Table 8
Comparison of Accuracy with and without Stacking Summary of techniques.

	Without Stacking	With Stackin
Dataset 1	0.7532	0.7635
Dataset 2	0.8889	0.9304
	Accuracy of Dataset 1	**Accuracy of Dataset 2**
Only with Logistic Regression	0.7403	0.8803
With Feature Selection	0.7532	0.8889
With Ensemble Technique – Max Voting	0.7783	0.9341
With Ensemble Technique – Stacking	0.7635	0.9304

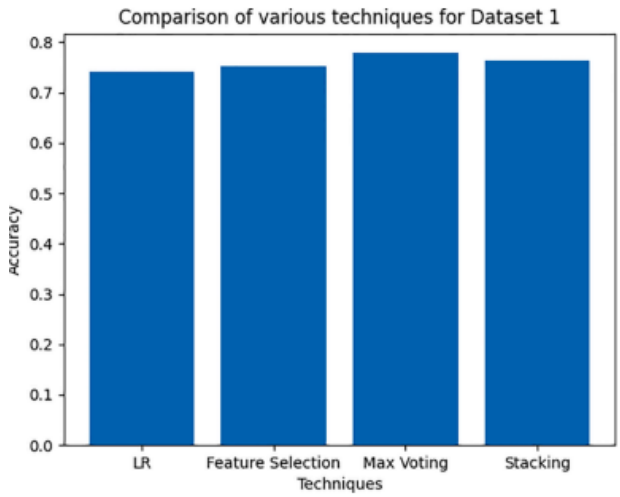

Fig. 6. Bar plot depicting accuracies of various techniques for Dataset 1

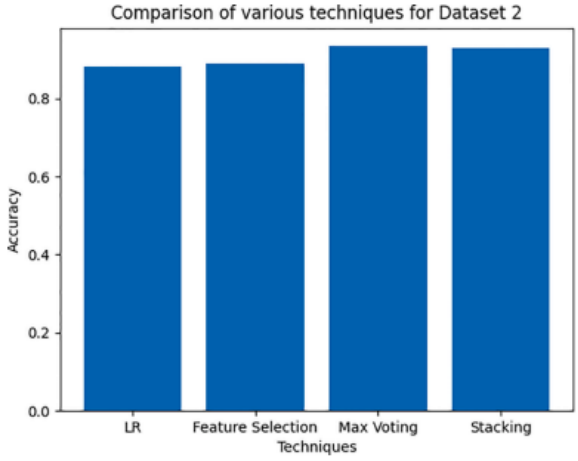

Fig. 7. Bar plot depicting accuracies of various techniques for Dataset 2

VI. CONCLUSION

It has been demonstrated that one of the effective techniques for creating prediction models is logistic regression. The method selected is just one of several variables that affect the model's accuracy. Preprocessing of data is one such element. In order to improve efficiency, redundant and null values must be removed. When characteristics diverge on a broad scale, normalizing the values also has a significant impact. As we have shown in this study, feature selection is crucial for improving accuracy and decreasing runtime. Combining several methods, as shown in ensemble approaches, contributes to the model's improved performance. Cross-validation is also crucial for increasing accuracy.

REFERENCES

[1] Rajendra, Priyanka, and Shahram Latifi. "Prediction of diabetes using logistic regression and ensemble techniques." Computer Methods and Programs in Biomedicine Update 1 (2021): 100032.
[2] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (Ijert) Volume 9 (2020).
[3] Qu, Kaiyang, Quan Zou, and Hua Shi. "Prediction of diabetic protein markers based on an ensemble method." Frontiers in Bioscience-Landmark 26.7 (2021): 207-221.
[4] Wadghiri, M. Z., et al. "Ensemble blood glucose prediction in diabetes mellitus: A review." Computers in Biology and Medicine 147 (2022): 105674.
[5] Wadghiri, M. Z., et al. "Ensemble blood glucose prediction in diabetes mellitus: A review." Computers in Biology and Medicine 147 (2022): 105674.
[6] Alehegn, Minyechil, Rahul Joshi, and Preeti Mulay. "Analysis and prediction of diabetes mellitus using machine learning algorithm." International Journal of Pure and Applied Mathematics 118.9 (2018): 871-878.
[7] Yadav, Dhyan Chandra, and Saurabh Pal. "An ensemble approach for classification and prediction of diabetes mellitus disease." Emerging Trends in Data Driven Computing and Communications: Proceedings of DDCIoT 2021. Springer Singapore, 2021.
[8] Saiti, Kyriaki, et al. "Ensemble methods in combination with compartment models for blood glucose level prediction in type 1 diabetes mellitus." Computer Methods and Programs in Biomedicine 196 (2020): 105628.
[9] Hasan, Mohammad Kamrul, et al. "An empirical model to predict the diabetic positive using stacked ensemble approach." Frontiers in Public Health 9 (2022): 792124.
[10] Kumar, P. Suresh, et al. "CatBoost ensemble approach for diabetes risk prediction at early stages." 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON). IEEE, 2021.
[11] Dutta, Aishwariya, et al. "Early prediction of diabetes using an ensemble of machine learning models." International Journal of Environmental Research and Public Health 19.19 (2022): 12378.
[12] Goyal, Priyanka, and Somil Jain. "Prediction of type-2 diabetes using classification and ensemble method approach." 2022 International Mobile and Embedded Technology Conference (MECON). IEEE, 2022.
[13] AC, Ramachandra, and Venkata Siva Reddy. "Bidirectional DC-DC Converter Circuits and Smart Control Algorithms: A Review." (2022).
[14] Kumari, Ashwini, et al. "Multilevel Home Security System using Arduino & GSM." Journal for Research 4 (2018).
[15] Viswanatha, V., et al. "Intelligent line follower robot using MSP430G2ET for industrial applications." Helix-The Scientific Explorer| Peer Reviewed Bimonthly International Journal 10.02 (2020): 232-237.

[16] Viswanatha, V., and R. Reddy. "Characterization of analog and digital control loops for bidirectional buck–boost converter using PID/PIDN algorithms." Journal of Electrical Systems and Information Technology 7.1 (2020): 1-25.
[17] Kibria, Hafsa Binte, et al. "An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI." Sensors 22.19 (2022): 7268.
[18] Viswanatha, V., R. K. Chandana, and A. C. Ramachandra. "IoT Based Smart Mirror Using Raspberry Pi 4 and YOLO Algorithm: A Novel Framework for Interactive Display." (2022).
[19] Laila, Umm E., et al. "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study." Sensors 22.14 (2022): 5247.
[20] Abdollahi, Jafar, and Babak Nouri-Moghaddam. "Hybrid stacked ensemble combined with genetic algorithms for Prediction of Diabetes." arXiv preprint arXiv:2103.08186 (2021).