A Machine learning approach to reduce the diabetes patient's readmission risk using a novel preprocessing technique

1st G G Rajput
Department of Computer Science,
Karnataka State Akkamahadevi
Women's University,
Vijayapura, Karnataka-586108
ggrajput@yahoo.co.in1.

2nd Ashvini Alashetty
Department of Computer Science,
Karnataka State Akkamahadevi
Women's University,
Vijayapura, Karnataka-586108
ashwinialashetty@gmail.com2

***Abstract*—High levels of glucose in the blood are the primary indicator of diabetes. Due to the fact that diabetes can cause several complications, many diabetics are hospitalized multiple times. The quality of treatment that is delivered and the overall cost of medical services are impacted when patients spend unnecessary time in the hospital. The healthcare systems inevitably capture data from clinical patient contacts; a machine learning based approach seems suited for addressing this issue. This paper includes one hundred thousand medical records for seventy thousand diabetic patients from 130 institutions in the United States (United States of America). Using a machine learning technique to process this massive dataset is time-consuming. Here, a unique preprocessing approach is demonstrated that reduces the 55 characteristics of the United States diabetes dataset to 21, with additional machine learning techniques used to predict early readmission. We employed machine learning (ML)-algorithms such as K-Nearest Neighbor (KNN), support vector machines (SVM), decision trees, random forests (RF), logistic regression (LR), gradient boosting (GB), and Gaussian naive bayes (GaussianNB) to evaluate the accuracy of a successful data engineering technique. Gradient Boosting, which beat the other machine learning methods, achieved 77.4% accuracy in the end.**

Keywords— Machine learning; diabetes; hospital readmission; pre-processing; accuracy

I. INTRODUCTION (*HEADING 1*)

The cost of diabetes treatment has become a major burden on the healthcare system in the United States (US). In 2017, this expenditure reached around $327 billion [1]. Ultimately, the hospital's capacity to forecast readmissions will aid in calculating and managing the quality of patient treatment. When a previously discharged patient is readmitted to the hospital within a certain time frame, this is known as a hospital readmission [2]. It is an indication of the treatment's efficacy and can have a significant influence on healthcare expenses. It may be difficult for hospitals to identify patients who are likely to require readmission within 30 days of release. Methods that forecast the likelihood of readmission and identify the circumstances that lead to readmission can be of great benefit to healthcare personnel. By employing such strategies, healthcare practitioners are able to tailor their interventions for high-risk patients and decrease readmission rates [3]. Rehospitalisation is indicative of the quality of care received, and healthcare costs are growing. Diabetes, like other chronic conditions, is connected with a higher risk of readmission to the hospital [4]. In this work, we explore several machine learning algorithms for assessing that diabetes patients would require hospital readmission. More than one hundred thousand diabetes patient records and 55 attributes, including duration of stay, use of insulin, and in-patient visits, were included in the data set utilized for this work. We employ a variety of pre-processing techniques and assess the performance of various models. The employed classifiers include of decision tree, random forest, SVM, k-neighbours, RF, GB, and GNB.

The rest of the paper is organized as follows: Section 2 goes on the methods for getting back into a hospital for diabetes patients. Section 3 introduces the new concept of the pre-processing system. The ML techniques employed are discussed in Section 4. Section 5 compares the ML model accuracy and section 6 summarizes the work carried out in this paper.

II. LITERATURE SURVEY

Ida et al. [5] aim to enhance accuracy by lowering the number of characteristics by a) contacting a physician and b) using feature selection filter techniques to algorithms created by the core module. Using the Naive Bayes classifier and electronic health records to analyze COPD patient data, Piyush Jain et al. [6] observed that parallel computing can cut processing times while retaining overall model performance when applied to COPD patient data. Utilized metrics included recall, precision, and cluster time. Wang et al. [7] collected two datasets from the Barnes Jewish Hospital's general hospital ward and operating room data in order to predict hospital readmission using deep learning.

A logistic regression model for diabetes classification was suggested by Qawqzeh et al. [8]. They used 459 patient records to train the model, then used another 128 records for testing and validation. Their recommended method had a

979-8-3503-9747-5/22/$31.00 ©2022 IEEE

92% accuracy rate and correctly diagnosed the absence of diabetes in 552 individuals. Pethunachiyar provided a method for classifying diabetes mellitus based on algorithms for machine learning [9]. He largely used a support vector machine equipped with a number of kernel functions and diabetes data from the UCI Machine Repository. His research showed that the SVM with a linear function outperformed neural networks, decision trees, and naive Bayes. However, the comparison to the current state of the art is absent, and the selection of parameters is not addressed.

Gupta et al. [10] classified diabetes using a support vector machine and a naive Bayes technique. PIMA's dataset on diabetes in India was utilised. In order to increase the accuracy of the model, k-fold cross-validation and a feature selection-based technique were also employed. Choubey et al. published an article comparing several diabetes categorization methods [11]. They employed both PIMA Indian data from the UCI Machine Learning Repository and a local diabetes dataset. Using AdaBoost, K-nearest neighbour regression, and radial basis function, they were able to determine whether or not people in both datasets had diabetes.

When developing their diabetes prediction system, Kumari et al. [12] relied on three popular supervised machine learning techniques. Evaluations were performed using data from the PIMA and breast cancer databases. Results from random forest, logistic regression, and naive Bayes were evaluated and contrasted with those from state-of-the-art single-method and ensemble approaches. This article by Hussain and Naaz [13] provides a comprehensive analysis of the machine learning techniques used for diabetes prediction from 2010 to 2019. They compared classic supervised machine learning models to modern neural network-based methods to figure out which was more accurate and productive. Naive Bayes and random forest were shown to be superior than all other algorithms when measured with the Matthews correlation coefficient.

III. A Novel Pre-Processing Methodology

Real-world medical data is noisy, inconsistent, and incomplete. Thus, data must be efficiently pre-processed and prepared for predictive modelling before developing the prediction model. Feature selection has been demonstrated to be an effective and efficient data pre-processing technique for preparing data for various machine learning models.

A. Data Preparation

The information for this research came from the machine learning repository at UCI. The collection includes the medical records of 101766 diabetic patients from 130 U.S. institutions (United States of America). The data collection report from the various US cities is shown in Fig.1.

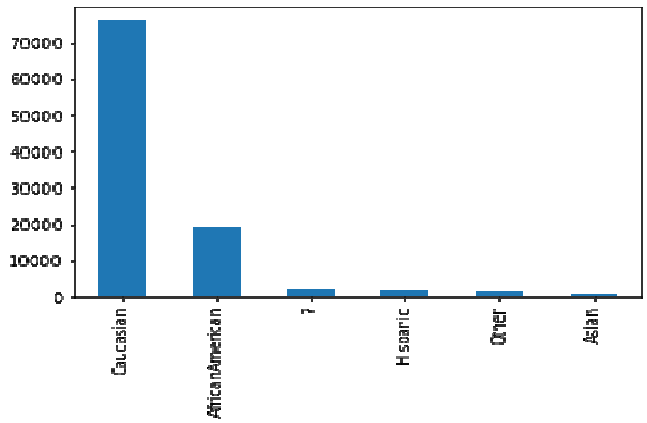

Fig. 1. Data collection report from the different cities

A question mark (?) is used in the dataset to indicate that there are some missing values that should be omitted from the dataset.

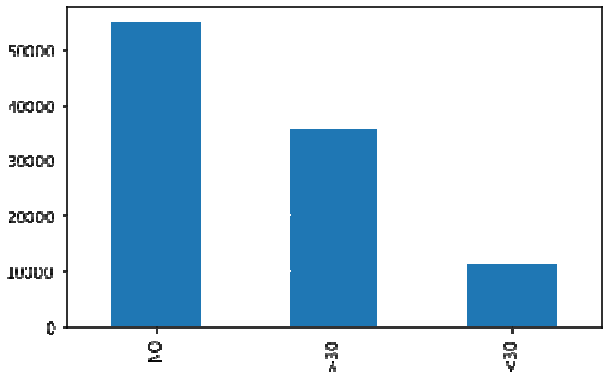

Fig. 2. Data prediction using Readmitted

According to Fig.2. Over 35545 patients were readmitted after 30 days, 11357 patients were readmitted before 30 days, and 54864 patients were not admitted to the hospital. The "discharge disposition id" column, which indicates the patient's post-hospitalization location, is another important piece of data. The mapping of the IDs is evident. According to the csv file given by UCI, numbers 11, 13, 14, 19, 20, and 21 are associated with death. These samples cannot be readmitted, thus they must be eliminated from the prediction model.

B. Feature Engineering

For each segment, new variables can be added to the data frame, and we can keep track of which columns to include in the prediction model. The numerical and categorical characteristics data's are present. This data collection's blank spots were supplemented by a question mark. In its place, let's use a "nan" representation. The addressing outliers in this dataset is unfeasible, and these characteristics may be left as-is. Categorical variables include non-numerical information such as race and gender. One-hot encoding is a technique that may be used to convert this non-numerical data into variables in the simplest way possible. Figure 3 demonstrates that the payer code, medical

specialty, and weight are missing more than fifty percent of the data, thus it is preferable to exclude these items.

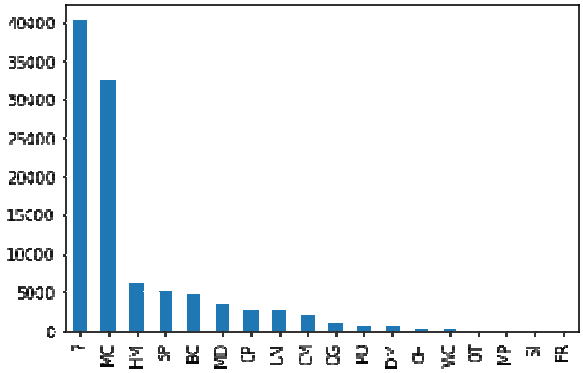

Fig.3. Payer code value

For each unique column value, we create a new column using one-hot encoding. The value of the column is 1 if the sample includes the specified value; otherwise, it is 0. We may include the columns for one-hot encoding in the data frame using the contact function. Axis = 1 must be used to indicate the addition of columns. The final two columns we want to include are age and weight. Since these values possess a natural order, it may be prudent to convert them to ordered numerical data. Let's convert the numbers in 0 to 90 via 10-point scales is shown in Fig.3.

Fig.4. Patients with age for each unique column value, we create a new column using one-hot encoding. The value of the column is 1 if the sample includes the specified value; otherwise, it is 0. We may include the columns for one-hot encoding in the data frame using the contact function. Axis = 1 must be used to indicate the addition of columns. The final two columns we want to include are age and weight. Since these values possess a natural order, it may be prudent to convert them to ordered numerical data. Let's convert the numbers in 0 to 90 via 10-point scales is shown in figure 4.

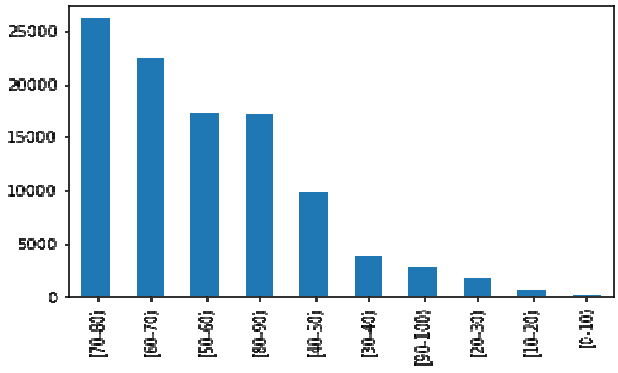

Fig.4. Patients with age

All 23 columns pertain to the quantity of medications supplied to a hospitalized patient. Given that the 23 Drugs were used to generate the treatments column, they will be deleted. Now, the number of data decreases to 26,820 with 24 columns.

C. Feature Identification

Further, we would like to eliminate columns "diag 1", "diag 2", and "diag 3" because they possess information on the codes for the numerous therapies delivered to the patient. They do not improve the effectiveness of the therapy. After deleting these columns, we were left with 21 columns containing 26820 data. The output variable is categorical, but both the discrete quantitative and categorical input variables are discrete. Due to the fact that we are integrating discrete quantitative data with categorical ones, we cannot do general correlation tests. The Chi-Square Test of Independence is used to assess whether or not the variables are correlated. The chi-square independence test reveals the usefulness of the twenty-one characteristics mentioned here.

- admission_type_id
- age
- discharge_disposition_id
- admission_source_id
- time_in_hospital
- num_procedures
- num_medications
- number_outpatient
- number_inpatient
- number_diagnoses
- race_AfricanAmerican
- race_Caucasian
- race_Hispanic
- gender_Female
- gender_Male
- max_glu_serum_>300
- max_glu_serum_None
- max_glu_serum_Norm
- A1Cresult_>7
- change_Ch
- change_No

IV. ML MODEL IMPLEMENTATIONS

K-NN, Decision Tree, Random Forest, Logistic Regression, Gradient Boosting (GB), Gaussian Naive Bayes (GaussianNB), and SVM are compared to the preprocessed dataset. Training and testing data, which comprised 70% and 30% of the total, were separated. Figure 5 depicts the suggested model flow. Training and testing datasets are (18774, 21) and (8046, 21). Our models should be more accurate than the basic model, which is 54%.

A. KNN

KNN is among the most fundamental models in machine learning. After the model analyses the k nearest data points for a single sample point, the probability is obtained by counting the positive labels and dividing by K. Although this model is simple to use and comprehend, it is

sensitive to K and requires time to analyze if there are several training samples.

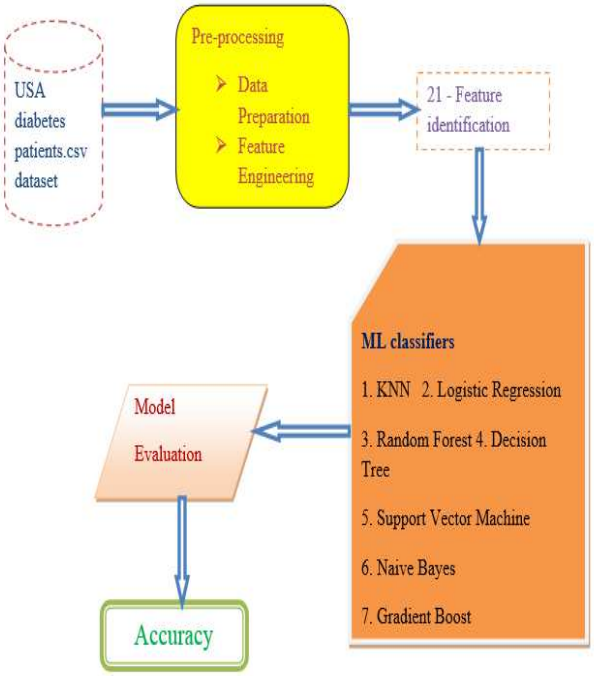

Fig.5. Proposed Methodology flow Diagram

B. *Logistic Regression*

Logistic regression can fit a linear positive-negative sample choice. A sigmoid function processes this linear function to calculate the likelihood of the positive class. Logistic regression works effectively for linearly separated features.

C. *Decision tree*

Decision trees are the simplest tree-based algorithm. The tree's last leaf's positive sample proportion determines the forecast. Machine learning determines which variable and threshold to use at each split. Tree-based techniques may detect non-linear effects and make no data structure assumptions.

D. *Support Vector Machine (SVM)*

SVM [14] outperforms decision trees and logistic regression. A kernel technique handles nonlinear input spaces. The dataset is accompanied with 21 attributes. After the dataset is loaded, it is separated into training and test sets so that the effectiveness of the model can be assessed.

E. *Random forest*

Decision trees [15] often over fit by remembering training data. To avoid over fitting, random forests were constructed. Random forest models build many trees with varying outcomes. A random sample collection and feature distribution bind forest trees to decor. Random forests generally outperform decision trees because they generalize better.

F. *Gradient boosting*

Boosting is an alternative strategy for improving decision trees. Using this strategy, we construct a large number of shallow trees to repair the errors created by previously trained trees. Gradient boosting classifier refers to a model that combines this technique with a gradient descent method.

G. *Naive Bayes*

Naive Bayes is a distinct machine learning model that is applied sparingly. In Naive Bayes, the Bayes Rule is employed to compute the probability. This model is "naive" since it incorrectly assumes that each characteristic is independent.

V. Result and Discussion

Python and Scikit-learn were used to create the ML models described above.

TABLE I. COMPARISON OF TRAINING ACCURACY AND TESTING ACCURACY OF ML MODELS

Models	Training Accuracy	Testing Accuracy
Logistic Regression	0.774262	0.767711
KNN	0.797859	0.673130
Decision Tree	0.996591	0.698235
Random Forest	0.996591	0.754412
Tuned Decision Tree	0.777565	0.767214
Tuned KNN	0.745339	0.697117
Tuned Random Forest	0.996484	0.756152

The models were only modestly enhanced by the hyper-parameter adjustments shown in Table 1. The accuracy of predictions served as the key evaluative factor in this paper. We determined that the training accuracy was 79% and the testing accuracy was 66% after applying KNN to a newly processed dataset. We were able to achieve a training accuracy of 77 percent and a testing accuracy of 76.8 percent using LR. We discovered that the training accuracy of a decision tree was 99 percent and the testing accuracy was 69.8 percent. The random forest should have a training accuracy of 99 percent and a testing accuracy of 75.4%.

Accuracy can be determined by applying equation 1 & 2. Accuracy here refers to the total success rate of the algorithm.

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

Figure 6 shows the confusion matrix interpreting true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)

Then, TPR and FPR are defined as follows:
TPR = TP / TP + FN
FPR = FP / FP + TN

Figure 7 shows that the projected model has a lower false negative value of 579, a true positive value of 9366, a false positive value of 5309, and a false positive value of 11566.

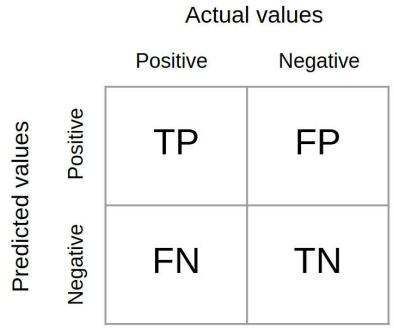

Fig. 6. Confusion matrix

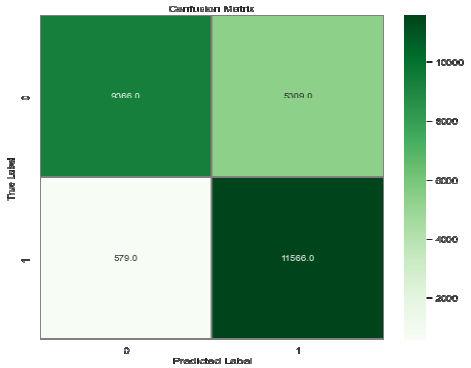

Fig. 7. Confusion matrix of the predicted model

Figure 8 demonstrates that, when compared to other ML models, Gradient Boost provides the greatest accuracy of 77.49%. The KNN, SVM, GNB, LR, Decision tree, and Random forest have an accuracy of 74.04%, 77.1%, 76%, 77.1%, 70.4%, and 72.8%, respectively. In table 2, SVM and LR are closer to GB, although GB is quantitatively superior to the other ML models.

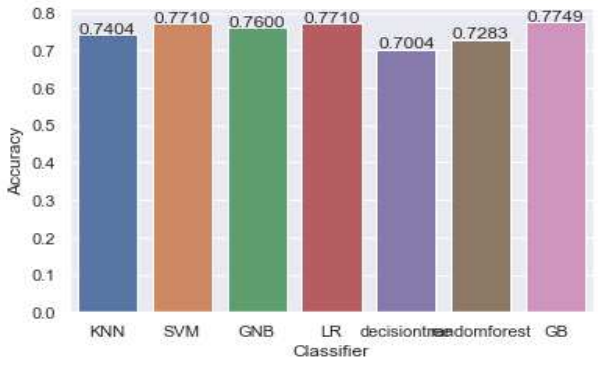

Fig. 8. Accuracy comparison with different ML models

TABLE II. ACCURACY COMPARISON WITH PROPOSED ML MODELS

Model Name	Accuracy %
KNN	74.03
SVM	77.10
GNB	76.00
LR	77.10
Decision tree	70.03
Random forest	72.82
GB	77.49

TABLE III. PERFORMANCE COMPARISONS WITH DIFFERENT DATASETS USING ML MODELS

Authors	Algorithm	Dataset	Total number of attributes	Accuracy%
Sarwar et al[16]	Six ML algorithms are used to predict disease	PIDD data set	50	SVM and KNN Th 72%
Sneha et al [17]	RF, SVM, NB, DT, KNN	USA diabetes dataset	50	Navie bays algorithm 72.3%
Deepti et al [18]	NB, SVM	PIDD data set	50	Navie Bays algorithm 76.3%
Proposed work	SVM,LR,GB,GNB, RF,DT,KNN	USA diabetes dataset	21	Gradient Boost 77.4%

Table 3 compares the different datasets used to estimate the accuracy of diabetes patients' hospital readmission. In our innovative preprocessing method, the number of attributes was decreased to 21, and a gradient boost machine learning model attained an accuracy of 77.4%.

VI. CONCLUSION

We created and evaluated ML-based readmission prediction algorithms for the 26820 diabetic patient records with 21 attributes investigated in order to predict the readmission risks of diabetic patients. The seven machine learning algorithms utilized by this solution are for predictive analytics. The major objective is to create diabetes detection classification systems. In this instance of readmission prediction, it is evident that lowering the amount of characteristics considerably enhanced the accuracy of the machine learning models. We constructed a machine learning model that can identify the diabetic patients with the highest risk of readmission within 30 days. The most efficient model consisted of a gradient boosting classifier than other ML models. The model performed 1.5 times better than patient selection and was able to detect 77.4% of readmissions.

REFERENCES

[1] Alamer AA, Patanwala AE, Aldayyen AM and Fazel MT, "Validation and comparison of two 30-day re-admission prediction models in patients with diabetes: *Endocr Prac.* Vol. 25, no. 11, 2019.

[2] Soh JGS, Wong WP, Mukhopadhyay A*, "*Predictors of 30th day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with metaanalysis" *BMJ Open Diabetes Research and Care* 2020;**8:**e001227.

[3] J. C. Ramírez and D. Herrera, "Prediction of diabetic patient readmission using machine learning," *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, 2019, pp. 1- 4, DOI: 10.1109/ColCACI.2019.8781796.

[4] Ida B. Seraphim, Varshita Ravi, Anchita Rajagopal "Prediction of Diabetes Readmission using Machine Learning", International Journal of Advanced Science and Technology, Vol. 29, No. 6, 2020, pp. 42Th49

[5] Piyush Jain, Ankur Agarwal, Ravi Behara and Christopher Baechle, *"HPCC based framework for COPD readmission risk analysis", Springer* (2019), 6:26

[6] Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, Alexander Kronzer, "Predicting Hospital Readmission via CostThSensitive Deep Learning", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (2018),* Volume: 15 Issue: 6, pp. 1968 – 1978.

[7] Y. K. Qawqzeh, A. S. Bajahzar, M. Jemmali, M. M. Otoom, and A. thalamus, "Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling," *BioMed Research International*, vol. 2020, Article ID 3764653, 6 pages, 2020.

[8] G. A. Pethunachiyar, "Classification of diabetes patients using kernel-based support vector machines," in *Proceeding of the 2020 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4, IEEE, Coimbatore, India, January 2020.

[9] S. Gupta, H. K. Verma, and D. Bhardwaj, "Classification of diabetes using Naıve Bayes and support vector machine as a technique," *Operations Management and Systems Engineering*, Springer, Singapore, pp. 365–376, 2021.

[10] D. K. Choubey, M. Kumar, V. Shukla, S. Tripathi, and V. K. Dhandhania, "Comparative analysis of classification methods with PCA and LDA for diabetes," *Current Diabetes Reviews*, vol. 16, no. 8, pp. 833–850, 2020.

[11] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using a soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, 2021.

[12] A. Hussain and S. Naaz, "Prediction of diabetes mellitus: a comparative study of various machine learning models," in *Proceeding of the International Conference on Innovative Computing and Communications*, pp. 103–115, Springer, Delhi, India, January 2021.

[13] A. Mahabub, "A robust voting approach for diabetes prediction using traditional machine learning techniques," *SN Appl. Sci.*, vol. 1, no. 12, 2019.

[14] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," *1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc.*, no. 2, pp. 1–4, 2019.

[15] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," *2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021*, pp. 141–146, 2021.

[16] A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," *ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput.*, no. September, pp. 1–6, 2018.

[17] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, 2019.

[18] Deepti Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.