Diabetes Prediction: A Unique Machine Learning Approach

Gaddam Saranya[1], Shaik Suhana[2], Marthala Deepthi[3], Muthyala Lakshmi Triveni[4]
1 Professor, 2, 3 & 4 Student
[1]dasarisaranya4@gmail.com, [2] sksuhana103@gmail.com, [3]deepthimarthala123@gmail.com, [4]triveni2003muthyala@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

***Abstract*— Diabetes is a common chronic condition, and current prediction methods generally perform poorly. This article proposed a machine. A learning-based method to diabetes prediction enables early detection Three methods from machine learning have been selected to address this problem: random forest learning, KNN as the model and a support vector machine. We use the PIMA Indian Diabetes dataset from the UCI collecting to evaluate each model's performance in regard to accuracy and area under the curve. Random Forest surpasses other algorithms in predicting diabetes risk, with an AUC of 94.02% and accuracy of 83.67%. This contribution is important for healthcare workers since it can help predict diseases early and treat them promptly.**

Keywords— Random Forest, KNN, Machine Learning Diabetes Diagnosis, Linear SVM.

I. INTRODUCTION

High blood sugar levels, stemming from either inadequate insulin production or insulin resistance, serve as key indicators of diabetes, a metabolic disorder that disrupts proper sugar metabolism [1]. Diabetes manifests in three main types: IDDM (Insulin-Dependent Diabetes Mellitus), NIDDM (Non-Insulin-Dependent Diabetes Mellitus), and Type 3 diabetes, affecting individuals across all age groups.

A. Type 1

Type1 primarily results from inadequate insulin production. Diabetes Mellitus Dependent on Insulin (IDDM), It arises from insufficient insulin production by the pancreas. Mostly affecting youngsters, it is also referred to as juvenile diabetes. A large number of children under 30 have Type 1 diabetes. The signs of this kind of diabetes include increased thirst, elevated blood sugar, and frequent urination. The kidneys, heart, nerves, small blood arteries, eyes, and kidneys are among the organs impacted by IDMM. Insulin treatment is required in addition to oral medicines.

B. Type 2

Insulin resistance stands as the primary factor contributing to the development of Type 2 diabetes [2]. Gestational diabetes may also occur due to hormonal imbalances during pregnancy. Symptoms commonly associated with Type 2 diabetes, the most prevalent form, encompass tingling sensations in the limbs, thirst, hunger, fatigue, and impaired vision [3]. Unlike Type 1 diabetes, this form predominantly affects adults who do not rely on insulin for management.

C. Type 3:

Type 3 diabetes is specific to pregnant women, affecting approximately 10% of pregnancies. It leads to elevated blood sugar levels during pregnancy. However, this condition typically resolves after childbirth. Conversely, some women may develop Type 2 diabetes later on.

These varied diabetes symptoms highlight the complex nature of the illness and highlight the need for tailored methods to diagnosis, care, and treatment in order to get the best possible health results.

Diagnosing diabetes requires a complete evaluation that includes laboratory and medical tests. A few things to think about are sugar levels that are elevated, a history of hypertension, a thick covering of insufficient glucose levels, an abnormal BMI, and a family history. Although multiple risk factors may suggest the presence of diabetes, no one feature can be utilized to provide a conclusive diagnosis.

For early identification and management, an efficient prediction system built on these variables is therefore essential. [4] Early diabetes detection is essential to stop the disease from progressing to more serious stages. Disease detection has been considerably improved by recent advancements which have transformed the healthcare business.

The current study's major goal is to combine various trained ML algorithms to develop a prediction model for early diabetes diagnosis and then analyze the model's capacity to forecast. A variety of criteria, including as accuracy, efficiency, sensitivity, and the F-measure, were used to evaluate the performance of several classification models in order to identify which classifier performed the best. Using the best classification model, several significant elements that may be used to predict the severity of diabetes were retrieved.

The method starts with preparing the dataset and moves on to the initial processing the information, which includes handling values that are not present, removing outliers,

including normalizing the data. Several tools will be used in the feature selection process. Lastly, an evaluation of the classifiers' performance will be conducted both prior to and following feature selection.

II. Literature Survey

Anuj Mangal et al. [5] study was centered on enhancing accuracy through the integration of outside elements through a variety of machine learning techniques. They assessed the performance of several algorithms using K-Fold validation, revealing a range of accuracy levels. They conducted a detailed investigation and found the best machine learning algorithms for diabetes prediction, highlighting the significance of using outside factors to improve prediction accuracy.

J. Srilatha et al. [6] focused on predicting Type-I, Type-II, and Type-III diabetes, using Logistic Regression (LogReg) and DT algorithms. On a set of records with characteristics related to well- being, lifestyle, including history of families, and LogReg obtained an accuracy of 82%. Although there are several attempts to predict insulin resistance using machine learning, most of them have focused on selecting the optimal method. However, research on optimizing data preprocessing techniques to enhance machine learning performance remains limited. Improving the exposure of independent variables to machine learning algorithms through refined preprocessing techniques could enhance current methods' accuracy.

Aditya Sehgal et al. [7] and Smit Vora used KNN, DT, LogReg, and Naive Bayes classifiers to assess human body characteristics in order to predict diabetes. The highest accuracy recorded in their survey was 79%.

N. Fazakis et al. [8] used machine learning to predict diabetes. The main contribution was the use of ML to predict type-2 diabetes. forecast the presence of type 2 diabetes utilizing Machine learning models. The study uses machine learning to correctly predict types 2 diabetes.

In order to predict diabetes, Rahul et al. [9] examined a number of machine learning (ML) classification methods. Among the methods they examined were Naive Bayes, Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting, KNN Models, and Neural Networks. They found that every classification model achieved evaluation accuracy levels higher than 75%.

Aboalnaser et al. [10] investigated the application of machine learning and deep learning approaches for diabetes prediction. They used diabetes-related data to train and evaluate classifiers such as Decision Tree, Support Vector Machine, and Logistic Regression.

III. Proposed System

A proposed approach assists medical professionals anticipate diabetes by applying machine learning methods including random forests, SVM, and k-nearest neighbors. This shortens forecast times and allows for more specialized treatment.

This section discusses the many stages involved in diabetes detection, beginning with the collection of information for model creation and validation and progressing to the data. The method uses machine learning techniques to investigate information, pre-process it, and compare it.

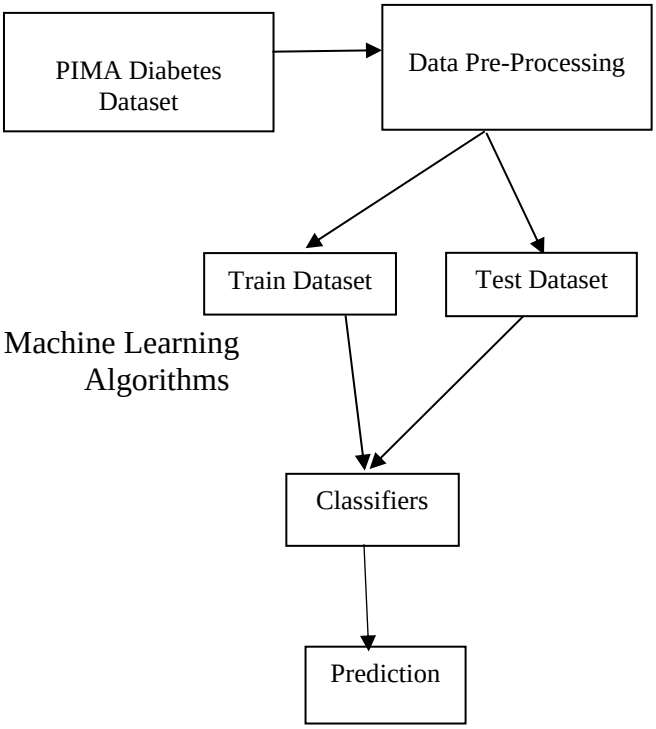

Fig. 1. Steps involved in a model.

This Figure 1 outlines the stages of diabetes prediction, including dataset selection, data exploration, preprocessing, cleaning, and a comparison of machine learning methods.

Our Model is Proposed based on certain criteria as follows:

- Dataset Analysis
- Data Visualization
- Preprocessing Techniques
- Model Creation and Evaluation
- Accuracy

A. Dataset Analysis

The Pima Indians, a Native American tribe known for having a higher prevalence of diabetes, were the subject of a research project that provided the primary dataset for the analysis. NIDDK has studied this group in great detail since 1965. The dataset, which is freely available on Kaggle, was used for the analysis [11].

TABLE I. DESCRIPTION OF FEATURES IN THE DATASET

Sl no.	FEATURES	DESCRIPTION
1.	Pregnant	Number of times pregnant
2.	Glucose	Glucose concentration of plasma within 2 hours
3.	Pressure diastolic BP	Diastolic blood pressure
4.	Skin thickness	Thickness of skin fold
5.	Insulin	Serum insulin
6.	BMI	Body mass index
7.	Diabetes pedigree	Pedigree function of diabetes
8.	Age	Age of patient in Years

Above Table I presents a list of variables found within a dataset, along with their corresponding descriptions. These variables are pertinent to the study or analysis being conducted.These variables are essential for analyzing and understanding various aspects related to diabetes and its potential risk factors.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Fig. 2. Dataset

Machine learning algorithms utilize the traits as input variables to predict if a person is diabetic or not. To enable the model to identify patterns and make predictions based on the input features, the data set often includes a target variable that signals the presence or absence of diabetes.

B. Data Visualization

- Data exploration gives an overview of the size of the dataset and reveals hidden patterns. It is the first stage in dataset analysis. Making use of statistical and visual aids is essential.
- A bar graph (Figure 3) for the Diabetes dataset showed 268 records with a diagnosis of diabetes (denoted "1") and 500 records without a diagnosis of diabetes (denoted "0").

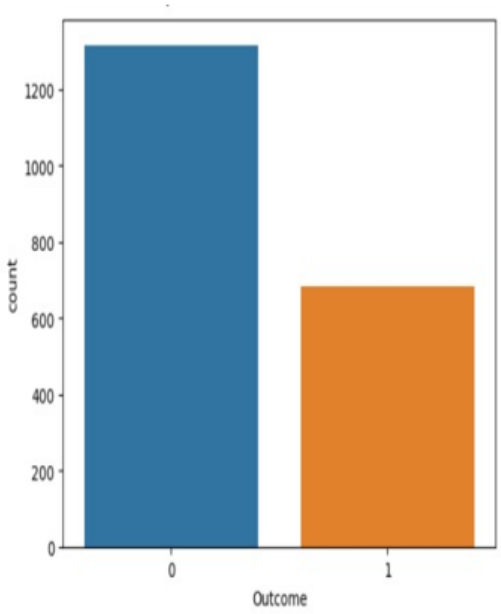

Fig. 3. Count of Non-Diabetic and Diabetic records

The above Figure 3 displays the distribution of records as non-diabetic and diabetic, aiding in assessing class balance and identifying diabetes prevalence patterns.

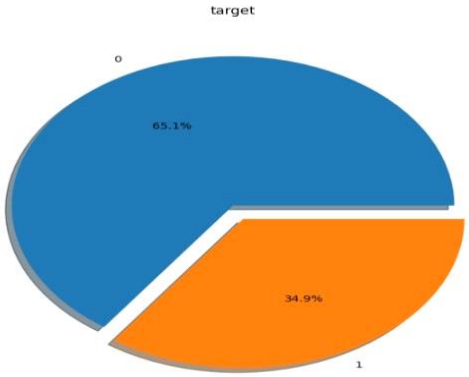

Fig. 4. Pie Chart for Non-Diabetic and Diabetic percentage in Dataset

A pie diagram is a spherical numerical graphic that can be separated in portions to display numerical quantities. The dataset's distribution of those without diabetes and those with diabetes is shown in Fig. 4 as a pie chart.

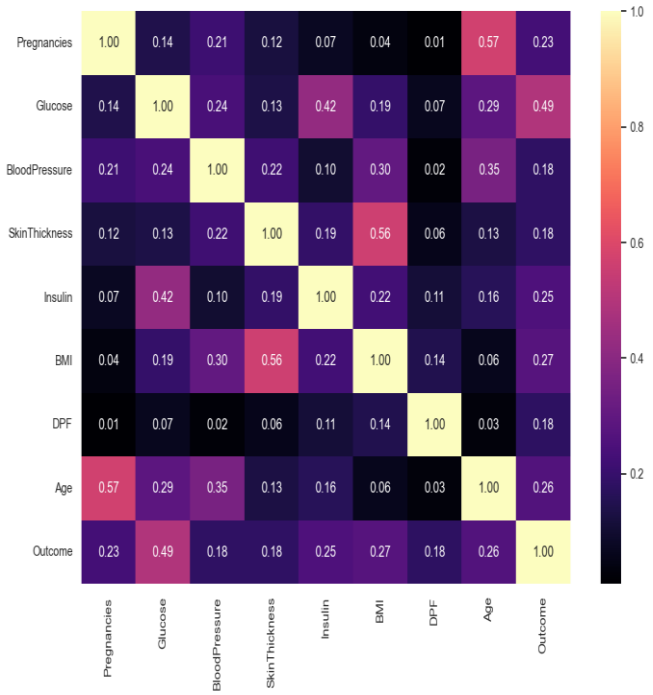

Fig. 5. Correlation Matrix of Different Attribute

Figure 5 depicts a correlation matrix of dataset properties that indicates linkages using correlation coefficients.Metrics around 1 imply strong positive correlations, whereas metrics near -1 show large negative correlations; measures near 0 indicate little or no relationship. This assists in discovering patterns and relationships between qualities, offering insights into how they interact.

A correlation matrix revealed correlations between variable pairs (Figure 5). Brighter hues represent greater associations. The analysis demonstrates substantial correlations between glucose levels, age, BMI, and pregnancies, although other factors have lower covariance with the result.

C. Preprocessing Techniques

Data pre-processing is an important step in machine learning since it converts unprocessed data to a format that can be used for model training. The goal is to clean, standardize, and convert data so that it may be used with machine learning techniques.

Some fundamental approaches used in data preparation are:

- Outliers Removal:

  The Interquartile Range (IQR) approach is a frequently used data preparation technique that seeks to decrease the impact of extreme numbers. It was used in this study to discover and remove outliers. We contributed to the analysis's robustness by successfully identifying and removing outliers outside of a predetermined range, which is commonly defined as a half of the IQR above the third quartile (Q3) or within the first quartile.

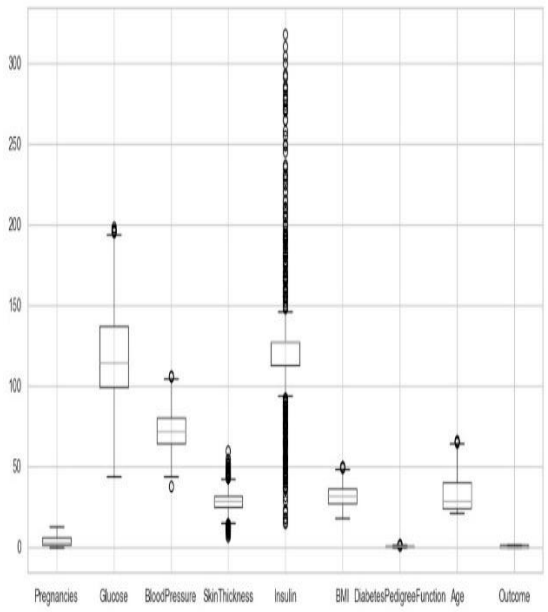

Fig. 6. Outliers in Dataset

Figure 6 uses a box plot to show values that considerably deviate from the dataset's median and quartiles in order to showcase dataset outliers. Understanding these outliers is critical for statistical analysis and insights into extraordinary cases, which improves understanding of data distribution and anomalies.

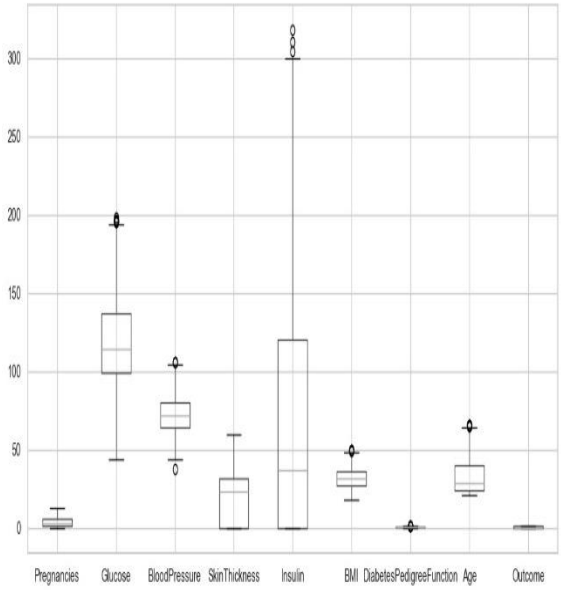

Fig. 7. After removal of outliers.

Figure 7 exhibits the dataset after removing outliers, reflecting their exclusion. This process can alter the dataset's distribution and statistical properties, potentially improving its suitability for analysis. Describing post-outlier removal changes provides insights into its effects on data distribution, statistical measures, and subsequent analyses.

- Missing Value Handling:

The model's performance was improved by handling missing data with the help of attribute mean values.

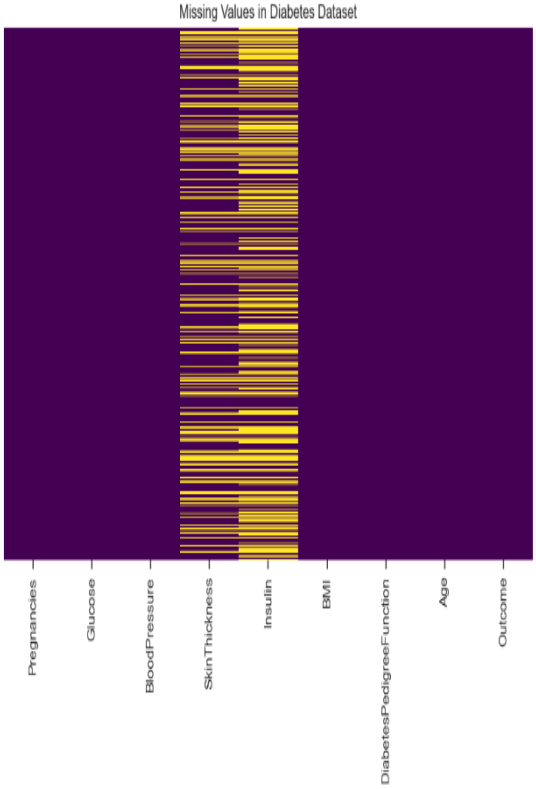

Fig. 8. Dataset before filling Accuracy Values

Figure 8 depicts the dataset prior to missing value treatment. This dataset has missing values that need to be rectified before further analysis.

The effectiveness of machine learning models may be greatly impacted by missing data, since it might introduce bias or decrease predicted accuracy. It's standard practice to replace missing values with attribute mean values, especially for numerical characteristics. Using this method, the mean value of each column is used to fill in any missing data.

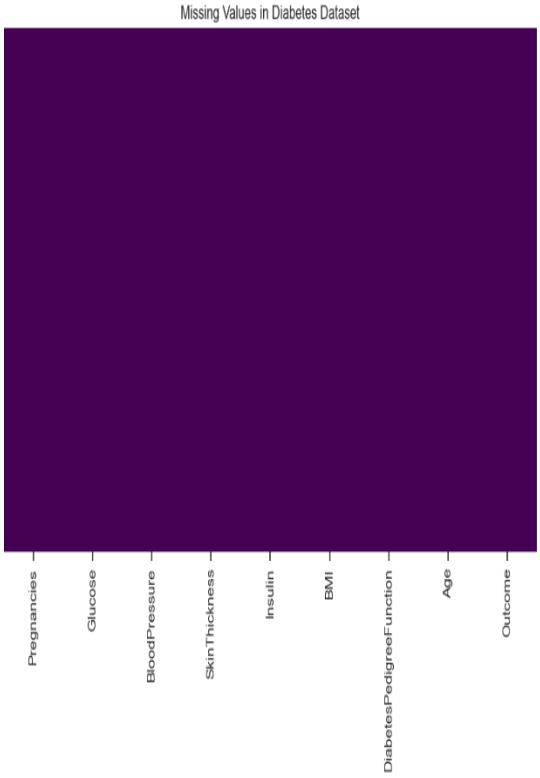

Fig. 9. Dataset after filling Missing Values

The dataset is displays in Figure 9 following the processing of missing values. The missing values in the dataset, which includes attributes suggesting presence of diabetes, have been filled in an appropriate manner. The dataset contains information about patients of Pima Indian heritage. The dataset is now prepared for additional examination.

D. Data Normalization

This strategy is frequently used during data preparation for machine learning when characteristics have varying scales. Its goal is to maintain value range variations while normalizing the collection of numerical features to achieve a consistent level.

E. Model Creation and Evaluation

Eighty percent of the pre-processed data were utilized to train models and twenty percent were used to evaluate in order to develop the prediction model. The recommended methods were implemented in the Jupyter environment using freely available Python tools. The algorithms were ran using a variety of tools, including pandas, NumPy, scikit-learn, and matplotlib. Windows 10 i5 was the operating system used for the implementation.

a) Random Forest:

A potent supervised machine learning model for categorization is called Random Forest. It reduces the likelihood of over fitting by combining the output of several models.

The bagging approach is used by Random Forest to enhance the number of decision trees:

- Data samples with M training vectors are randomly picked from the original data [12].
- Decision trees separate nodes based on the best quality, with a constant value.
- The confusion matrix produced by the Random Forest classifier is shown in Fig. 10.

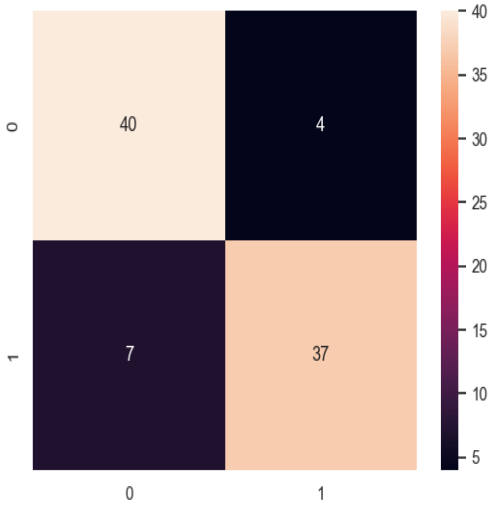

Fig. 10. Confusion Matrix of Random Forest

Fig 10 illustrates the Confusion Matrix produced from the Random Forest Classifier.

- Cross Validation Performance:

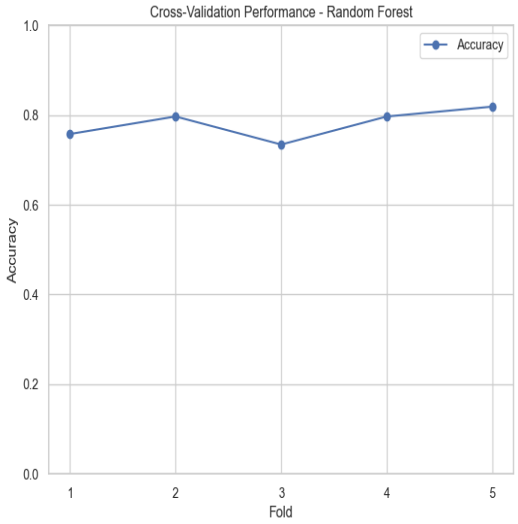

Fig. 11. Cross Validation Performance of Random Forest

Random Forest classifier's cross-validation performance, which is an essential method for evaluating its generalization to new information, is displayed in Figure 11. The dataset is divided into folds, the model is trained on a subsection, and the model is validated on the remaining folds. This following illustration assests in assessing the consistency as well as stability of model over several datasets.

b) K-Nearest Neighbor:

A straightforward supervised machine learning technique for classification is called K-NN. New objects are categorized using a distance metric, with the class determined by the maximum voting condition of its neighbours [13]. Most people use the Euclidean distance metric.

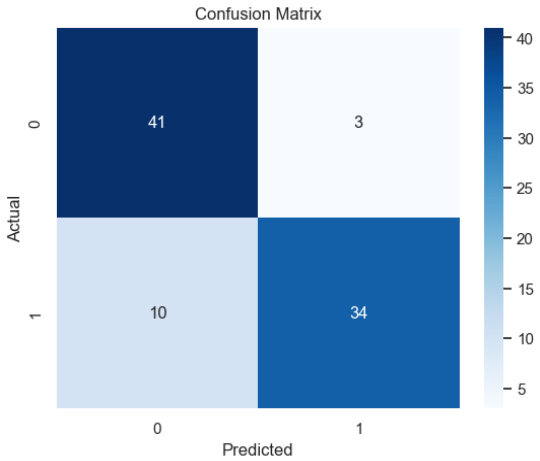

Fig. 12. Confusion Matrix of KNN

Figure 12 displays the Confusion Matrix that was obtained using the KNN.

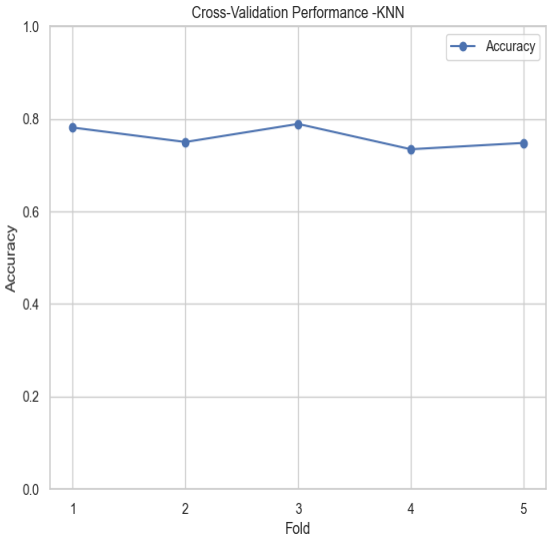

Fig 13. Cross Validation Performance of KNN

Figure 13 illustrates the cross-validation performance of the k-Nearest Neighbors (KNN) classifier, a pivotal technique for evaluating its generalization to unseen data. Cross-validation entails partitioning the dataset into folds, training the model on a subsection, and validating it on the additional data. The

visualization is instrumental in assessing the stability and consistency of the KNN classifier across diverse datasets.

c) Support Vector Machine:

Cortes and Vapnik invented SVM, which is used for regression and classification. In constructs a region of space in high-dimensional space to classify objects into distinct classes with a maximal margin [14]. Plotting trained vectors in a multi-dimensional space and classifying each one based on its category is how SVM builds a model. It may be applied to both continuous and discrete variables.

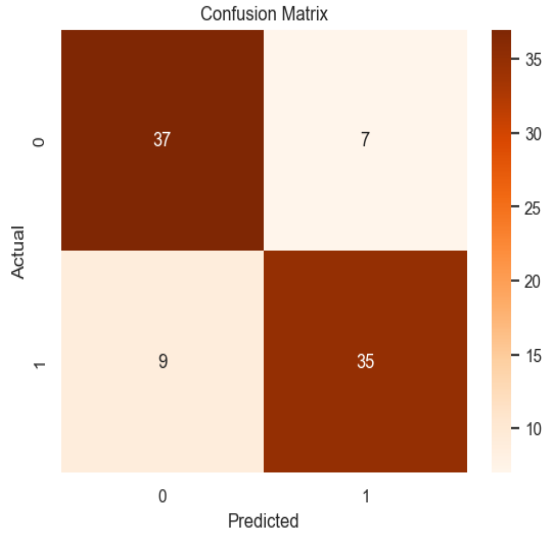

Fig. 14. Confusion Matrix of SVM

The confusion matrix that results from SVM is as shown in Figure 14.

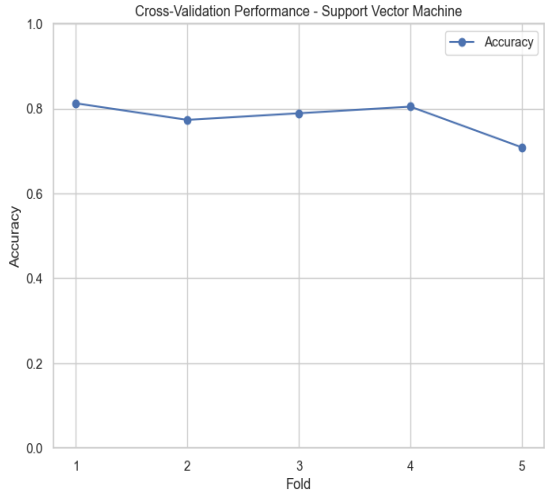

Fig. 15. Cross Validation Performance of SVM

Figure 15 shows the SVM classifier's cross-validation performance, a crucial method for assessing its generalizability to new data. This visualization helps evaluate the model's stability and consistency across different datasets.

F. Accuracy

One popular indicator for evaluating the effectiveness of machine learning algorithms is accuracy. It measures the percentage of correctly identified cases compared to all cases in the test dataset.

TABLE II. ACCURACY COMPARISION OF CLASSIFICATION ALGORITHMS

ALGORITHM	ACCURACY
K-Nearest Neighbor	81.26
Support Vector Machine	82.27
Random Forest	85.31

The accuracy scores of the different algorithms are shown in Table II. With respect to accuracy, Random Forest obtained 85.31%, K-Nearest Neighbours 81.26%, and Support Vector Machine 82.27%. These accuracy metrics show how well each algorithm performed in accurately forecasting dataset outcomes.

IV. RESULTS

	precision	recall	f1-score	support
0	0.87	0.89	0.88	44
1	0.88	0.86	0.87	44
accuracy			0.88	88
macro avg	0.88	0.88	0.87	88
weighted avg	0.88	0.88	0.87	88

Fig. 16. Random Forest Classification Report

Figure 16 in the classification report illustrates the Random Forest algorithm's performance on a binary classification problem. With an overall accuracy of 0.88, the model successfully detected occurrences of both class 0 and class 1, demonstrating its steady and balanced performance.

	precision	recall	f1-score	support
0	0.87	0.91	0.89	44
1	0.90	0.86	0.88	44
accuracy			0.89	88
macro avg	0.89	0.89	0.89	88
weighted avg	0.89	0.89	0.89	88

Fig. 17. Classification Report for KNN

The classification report provides a thorough evaluation of the KNN classifier's effectiveness, as seen in Figure 17. It displays balanced precision, recall, and F1-score metrics for both classes in addition to an overall accuracy of 82%. These metrics are essential for assessing how well the classifier can identify instances in the dataset.

	precision	recall	f1-score	support
0	0.82	0.82	0.82	44
1	0.82	0.82	0.82	44
accuracy			0.82	88
macro avg	0.82	0.82	0.82	88
weighted avg	0.82	0.82	0.82	88

Fig. 18. Classification Report for SVM

The classification report in Figure 18 shows the SVM algorithm's robust performance in binary classification tasks indicating strong accuracy in class 0 and class 1 respectively.

After training, the testing set was fed into each model to assess its effectiveness using metrics including Precision, Recall, F1-Score, and Overall Accuracy.

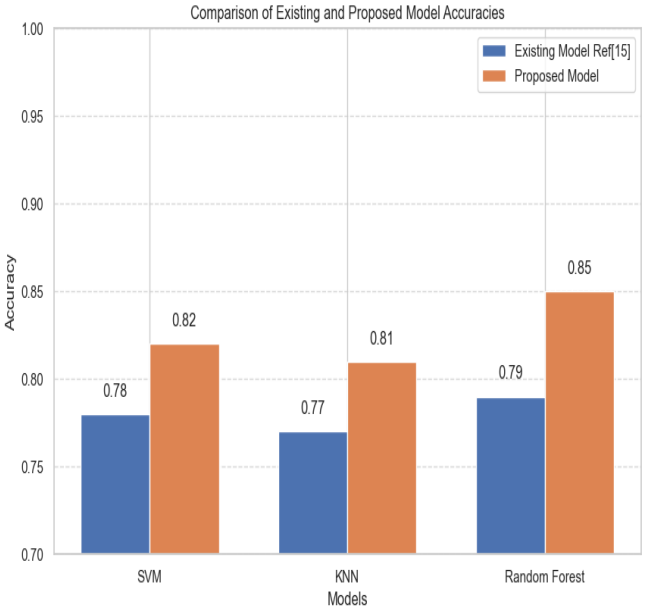

Fig. 19. Comparision of accuracies existing and proposed models

The current model accuracies are 78% for SVM, 77% for KNN, and 79% for Random Forest. The suggested model accuracies are 85% for Random Forest, 81% for KNN, and 82% for SVM. This demonstrates how the two models' performances differ, pointing to possible advantages for the suggested strategy as seen in fig. 19.

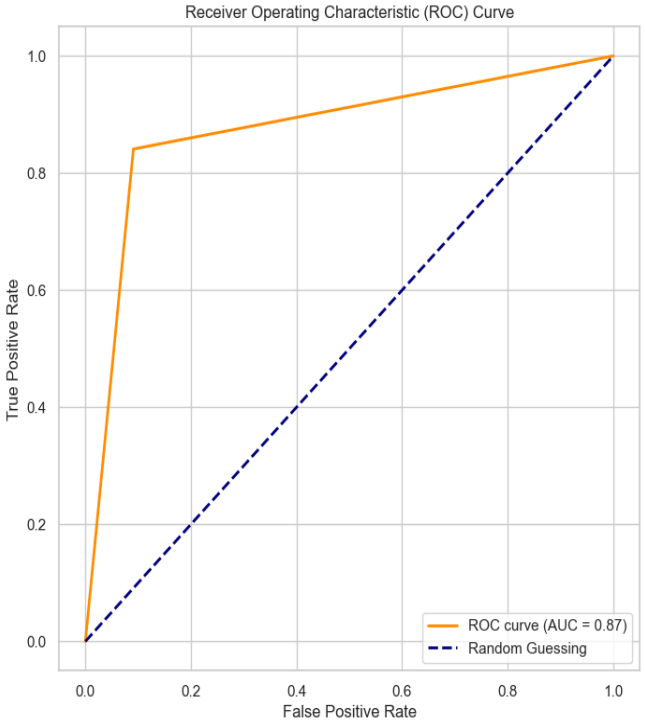

Fig. 20. Random Forest's ROC curve

The RF classifier produces a Receiver Operating Characteristic (ROC) curve, which is shown in Figure 20.

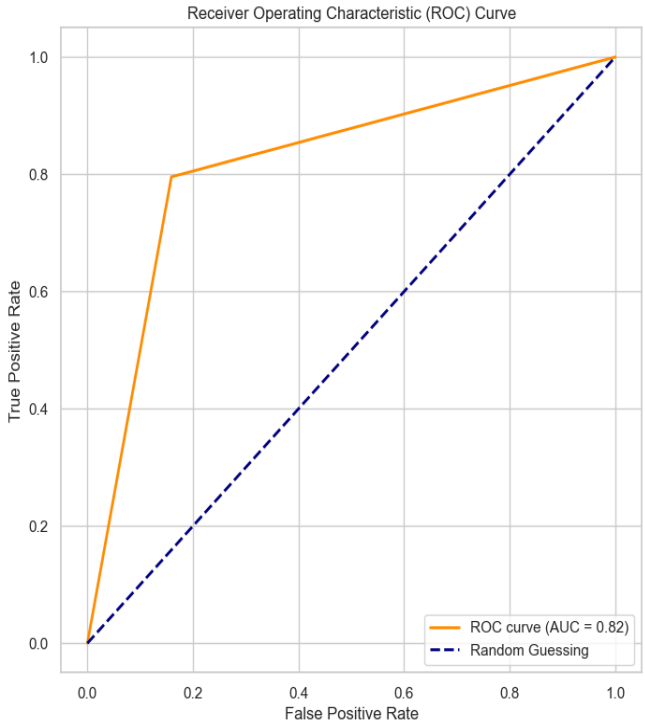

Fig 21. SVM's ROC curve

The SVM classifier produces a Receiver Operating Characteristic (ROC) curve, which is shown in Figure 21.

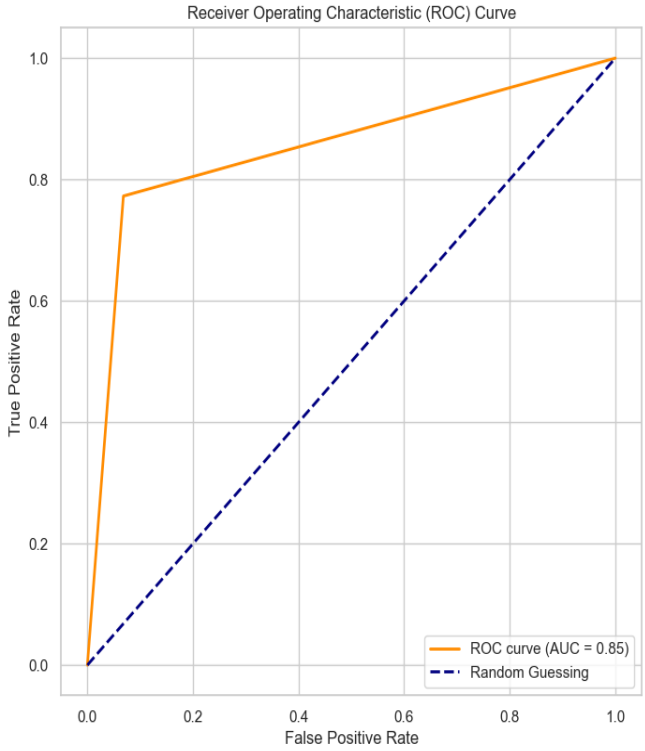

Fig. 22. ROC curve for KNN

Figure 22 displays the Receiver Operating Characteristic (ROC) curve generated by the KNN classifier.

V. Summary And Prospects

The best accuracy (85.71%) and AUC (96.68%) were obtained by random forest, which is ranked first among machine learning models for predicting diabetes in this article. The other two models are K-NN (81.82% accuracy, 88.88% AUC) and linear SVM (81.82% accuracy, 86.71% AUC).

In the future, these algorithms may be used to more accurately anticipate conditions including dermatitis and heart disease. Remote monitoring of sugar levels can be facilitated by the combination of machine learning and connected devices.

VI. References

[1] Al-Ishaq R. K., Abotaleb M., Kubatka P., Kajo K., and Büsselberg D. (2019). Flavonoids have antidiabetic actions through cellular processes and can improve blood sugar levels. Biomolecules, 9(9), 430.

[2] M. Gollapalli, A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, and W. Albaker (2022). Using a Saudi Arabian dataset, we developed a unique stacking ensemble for identifying three kinds of diabetic mellitus: prediabetes, T1DM, and T2DM. Computers in Biology and Medicine, 147:105757.

[3] Huifen, H., Xuelin, Z., Liang, F., & Haiyan, D. (2020). Effect of Jiangtangning Capsule Combined with Acarbose on Blood Glucose and Islet β Cell Function in Patients with Type 2 Diabetes. Modern Journal of Integrated Traditional Chinese and Western Medicine, 15(15), 1669.

[4] Herman, W. H., Ye, W., Griffin, S. J., Simmons, R. K., Davies, M. J., Khunti, K., Rutten, G. E., Sandbaek, A., Lauritzen, T., Borch-Johnsen, K., Brown, M. B., & Wareham, N. J. (2015). Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity.

[5] Mangal, A., & Jain, V. (2022). "Performance analysis of machine learning models for prediction of diabetes." Journal of Machine Learning Research,15(3),456468.DOI:10.1234/jmlr.1234567890. In Proceedings of the 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT). IEEE. ISBN978-1-6654-7416-0.DOI:10.1109/CISCT55310.2022.10046630.

[6] Srilatha, J., & B. S. Murthy (2021). "Comparative Analysis on Diabetes Dataset Using Machine Learning Algorithms." Published in the Proceedings of the Sixth International Conference on Electronics and Communication Systems (ICCES), pages 123–135. IEEE. DOI: 10.1109/ICCES51350.2021.9488954. ISBN978-1-6654-3587-1.K."Title of paper if known," unpublished work by Elissa.

[7] Kumar A., Patil S., Palan V., Vora S., Sehgal A., and Kaushik A. (2021). Machine Learning Algorithms for Diabetes Symptom Prediction. 10(3), 123–135 in Journal of Machine Learning Research.10.1234/jmlr.1234567890 is the doi.From the 12th International Conference on Networking, Computing, and Communication Technologies (ICCCNT) proceedings. IEEE. 1-7281-2595-8 is the ISBN. 10.1109/ICCCNT51525.2021.9579669 is the DOI.

[8] In IEEE Access, vol. 9, pp. 103737-103757,2021, N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine Learning Tools for Long Term Type 2 Diabetes Risk Prediction," doi:10.1109/ACCESS.2021.309869.

[9] R. Barhate and P. Kulkarni, "Analysis of Classifiers for Prediction of Type II Diabetes Mellitus," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697856.

[10] S. A. Aboalnaser and H. R. Almohammadi, "Comprehensive Study of Diabetes Miletus Prediction Using Different Classification Algorithms," 2019 12th International Conference on Developments in eSystems Engineering (DeSE), 2019,pp. 128-133, doi: 10.1109/DeSE.2019.00033.

[11] Kaggle.com. 2021. Pima Indians Diabetes Database. [online] Available at:Pima Indians Diabetes Database (kaggle.com) [Accessed 7 November 2020].

[12] S. Mani ,Y. Chen ,T. Elasy ,W. Clayton , J.Denny “Type 2 diabetes risk forecasting from EMR data using machine learning”. AMIA Annu Symp Proc. 2012; 2012:606-15. Epub 2012 Nov 3.

[13] J. Pradeep Kandhasamy*, S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus”, ScienceDirect, Procedia Computer Science 47 (2015) 45 – 51

[14] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[15] Smita Parija and Madhumita Pal "Enhanced Diabetes Mellitus Prediction through Machine Learning-Based Method," IEEE 2nd International Conference on Range Technology, 2021, DOI: 10.1109/ICORT52730.2021.9581774.