JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

A DIABETES PREDICTION DECISION SUPPORT SYSTEM USING MACHINE LEARNING

Muhammad Faisal[1], Imtiaz Ali Shah[2], Faiza Aurang Zeb[3]

Affiliations

[1] Department of Computer Science, University of Lahore Sargodha Campus, Pakistan

[2] Department of Computer Science, COMSATS University of Technology, Abbottabad, Pakistan

[3] Department of Computer Science, GPGC Mandian Abbottabad, Pakistan

Corresponding Author's Email

[3] faizaaurangzeb222@gmail.com

Abstract

Diabetes is a chronic metabolic disorder affecting millions worldwide and is a leading cause of serious health complications such as heart disease, kidney failure, and vision loss. Early detection and accurate risk prediction are vital for effective disease management and reduction of long-term health burdens. This study presents a robust diabetes prediction decision support system leveraging advanced machine learning techniques applied to a curated dataset obtained from Kaggle. The primary objective is to address the critical need for early and accurate diabetes diagnosis by developing a reliable and efficient prediction model. Through comprehensive data preprocessing, detailed feature engineering, and systematic model training, a Random Forest Classifier was developed and optimized using GridSearchCV. The resulting model achieved an impressive accuracy of 94.70%, demonstrating a precision of 0.98, recall of 0.96, and F1-score of 0.97 for non-diabetic cases (class 0), and a precision of 0.65, recall of 0.80, and F1-score of 0.72 for diabetic cases (class 1). These results underscore the model's high reliability in detecting non-diabetic instances and its reasonable effectiveness in identifying diabetic cases. The findings offer significant implications for clinical decision-making, enabling healthcare professionals to implement proactive interventions and personalized treatment strategies. Furthermore, the model provides a valuable foundation for future enhancements in medical data analysis and health informatics applications aimed at improving chronic disease management.

Keywords: Diabetes Prediction, Machine Learning, Random Forest Classifier, Health Informatics, Medical Data Analysis, Predictive

INTRODUCTION

Machine learning algorithms can evaluate big datasets of patient information, comprising medical history, genetics, lifestyle factors, and clinical measurements, to identify patterns and risk factors associated with diabetes. This enables healthcare professionals to assess an individual's risk of developing diabetes and intervene with preventive measures [1]. Diabetes refers to a group of metabolic disorders considered by elevated blood sugar stages persisting for an extended duration. Common indicators of high blood sugar encompass frequent urination, heightened thirst, and increased appetite. When left unmanaged, diabetes can give rise to a range of complications. Immediate complications may involve conditions like diabetic ketoacidosis, hyperosmolar hyperglycemic state, or even mortality. Furthermore, enduring and severe complications encompass ailments such as cardiovascular disorders, strokes, chronic kidney dysfunction, foot ulcers, and ocular impairment [2]. To determine whether the patient has diabetes centered on the features that we will provide to our machine learning model, and to do so, we will be using the Pima Indians Diabetes Database.

Machine learning has made a significant impact on the prediction and management of diabetes in the real world. In case of early detection and risk management, when looking for correlations with diabetes and risk factors, machine learning algorithms can sift through large databases of patients' sensitive data. This information can be patient's medical history, family history, patient's behavior [3], and some clinical tests results. It allows the care health provider to identify the precursors of the disease early and treat them to prevent the occurrence of the diseases. Personally, data given discovered medicine was discovered that through machine learning, it was possible to develop accurate models that considered a person's individual biography in the prediction and diagnosis of diabetes. Such models can help in the delivery of precise advice and instructions for alterations of the lifestyle, constant tracking, and interventions considering the person's risk factors.

To monitor the continuous glucose data, it is possible for machine learning algorithms to use live quantitative data from the CGM devices used by the diabetics [2]. Such algorithms can identify patterns and changes in glucose level that inform the patient and healthcare providers on the need to change medication, diet or insulin intake schedule. Furthermore, for needy diabetics, beneficiaries of insulin, information including but not limited to food intake, physical activities, glucose level, and the insulin, machine learning algorithms can quickly calculate the insulin dosage and the best time [3]. This can in turn help in improved glucose regulation and hence the chances of developing hypoglycemia are minimized.

Today, most applications and websites for mobile devices apply machine learning algorithms for diet tracking, exercising, and glucose levels monitoring. Some of these applications can help in viewing effects of certain decisions on the level of blood sugar and suggest ways through which it can be brought to the normal levels. Additionally, there is the ability to use the machine learning systems to support its recommendation system for clinical healthcare. These systems can also work on patient data to help in the diagnosis of the patient, in treatment planning as well as in management of the patient's medications [4]. There are various effective cases of utilizing machine learning in pharmaceutical research including [4] the search for drug candidates and improving the diabetic therapy strategy. It can manage the generation of more effective and personalized treatments. Also, Machine learning is used for remote diabetes patients' supervision with the help of wearable devices and telemedicine [5]. These tools allow the respective healthcare providers to monitor the health status of the respective patients and aid them as and when required. Machine learning used in this process assists the researchers to analyze the big data sets and identify new trends that can be associated with diabetes' risk factors, progression or effect of treatment control. They also help in increasing the knowledge of diabetes and directing the future research in this field [6]. Population-level data can also be analyzed using machine learning algorithms in the context of diabetes; that is, for the purpose of discovering patterns regarding the prevalence, risks, and prognoses of the condition. Such information is useful in the determination of public health measures as well as allocation of financial and human resources [7].

Finally, Machine Learning (ML), is used to assess medical histories of large population looking for elements that can expose people to diabetes and help interventions are taken in a way were diagnosis was made earlier than regular detection methods. Personalized recommendations done through machine learning enabled continuous glucose monitors with mobile apps that help in blood sugar control. Machine learning in drug development to enhance diabetes therapy and personalized treatments. Timely healthcare support is provided through remote monitoring and telemedicine that are also ML based. ML reveals new trends of diabetes research, that light the journey for future researchers; and population-based insight moves public-health strategy on resource allocation.

Contributions of the work are given below:

- Careful data preprocessing, including duplicate removal, missing values, and feature scaling, ensures data quality.

- Advanced feature engineering extracts useful facts from the dataset to improve model prediction.
- The durable and versatile Random Forest Classifier model was chosen for complex data patterns.
- Hyperparameter tuning using GridSearchCV optimizes model performance and selects the best parameters for predicted accuracy.

An introduction provides the background and overview of diabetic prediction, and the article is organized to flow smoothly. The literature review summarizes previous studies to support the research. The methodology section describes data collecting, statistical analysis, and machine learning algorithms. The results present the analysis's findings, while the discussion section interprets and examines their consequences. The conclusion summarizes the study's findings and advises future research to ensure complete understanding.

Related Work

The lasting condition known as diabetes (the disease that affects millions of people around the world) is one where increased sugar level can increase in the blood. This initiates either of these two situations: either the body does not have enough insulin or is not a good use of it. On the other hand, it is a very serious public health problem because it is related to several seriously dangerous disorders like cardio disease, kidney failure, and nerve damage.

In different studies, diagnosis, and prevention of T2DM by employing data mining was shown to be one of the current approaches. They are presented in multiple studies aiming to tackle the problem in several ways. Yang et al. [1] suggested a promising approach with the main objective of enhancing the accuracy of the prediction models combined with having the models generalized over several datasets in their study. Both a more advanced version of the k-means algorithm and logistic regression were adopted into their model with it being further enhanced through the addition of a pre-processing layer. Islam et al. [2] employed the Sylhet Diabetes Hospital surveys conducted in Bangladesh dataset aiming at forecasting the possibility of having diabetes among these people. The researchers ran the models Naive Bayes, Logistic Regression, and Random Forest on the dataset to achieve this objective. A similar method to that of Woldemichael and Menaire [3], who used data mining algorithms like the backward propagation algorithm, was adopted to determine the patient rate diabetes in the future. The stages of the diabetic complications model development by Fiarni et al. for diabetes in Indonesia involved applying k-means clustering and Naive Bayes Tree classification [4]. The software that Aldallal et al. [5] developed as a part of their project to forecast the beginning of chronic diseases was evaluated on the data of patients from the Hospital of Bahrain Defense Force [HBDH] once it was developed.

The machine learning methods applied in the studies by Khan et al. [6] as well as Kavakiotis et al. [7] were employed to establish computer-based prediction systems for glycemic management and diabetes research. Kumar and colleagues [8] devised a model aimed at assessing the accuracy of heart disease risk prediction in patients with diabetes while Mahesh et al. [9] applied a machine learning algorithm in their work. K-nearest neighbors and logistic regression are two algorithms among the predictive techniques that Oza and Bokhare used in their study. The Scientists Anil et al. [11] scrutinized the diabetes mellitus prediction using a different analysis method. Paisanwarakiat et al. [12] proposed the combination of logistic regressions and decision trees with the approximation of the data through the Naive Bayes (Ihab and Reddy, 2020). Besides this, Arumugam et al. [13], employed three machine learning practices incorporating Random Forest, K-nearest neighbor, and Support Vector Machine to come up with a classifier which had the capability of predicting whether a given person was at risk or not of becoming diabetic. The authors condemned that the results of the support vector methodology were sure. Finally, Abdollahi and Moghaddam [14] utilized the data on experimental and real Indians for the same and lead the genetic algorithm-based ensemble training strategy to head with the impacts of diabetes.

METHODOLOGY

The creation of a reliable diabetes prediction model with the representation of different stages is the objective of this collaborative process of ours. During the process of data collection, data cleaning, variable selection, model training, model evaluation, and model deployment, you are going to pass these phases. Each of the steps is critical and is incorporated to make sure that the prediction model is as true and accurate as possible.

1.1. Dataset Selection

The dataset of "Diabetes Prediction Dataset" is downloaded from Kaggle [15]. These tableau files contain 100,000 records of patients' age, gender, BMI, family history of diabetes, blood pressure, and blood glucose levels. Through in-depth examination of the risk associated with diabetes, we try to find the hidden patterns and trends that can lead to future studies on the given topic. This study investigates how these variables which act together cause and worsen diabetes. With its important role in health care, knowledge about it influences the kind of treatment and prognosis.

1.2. Data Preprocessing

First, we looked for null using IsNull(). Next, we will use sum() to find the total of all missing values and eventually conclude the total. Then we know that there are no missing values, but it is not the case. In the present dataset, all NAs are filled with 0 that is not good for reality as you can see in *Figure 1*. Once again for that, we will replace 0 and begin the imputation. After incorporating this approach, we can always impute it later to preserve the integrity of the dataset and figure out a better Imputation mechanism.

```
Output:

Pregnancies                   0
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

Figure 1: Data preprocessing

1.2.1. Attribute Description

The values of the medical predictor factors are given in the dataset near the single Outcome variable; the Outcome variable towards this dataset contains a wide range of parameters as predictor variables, and few of them are Age, number of Pregnancies, Body Mass Index (BMI), Insulin [8], etc.

Table 1: Attributes of Dataset

Attribute	Information
Pregnancies	The total number of pregnancies
Blood Pressure	blood pressure (diastolic)
Thickness of skin	Skin fold thickness of triceps
Glucose	Plasma glucose concentration during a two-hour oral glucose tolerance test.
Insulin	Serum insulin (mu U/ml) 2-hours
H	Body mass index (weight in kg/ (height in m) ^2)
Diabetes Pedigree Function	pedigree function of diabetes
Age	Age (years)
Outcome	Class variable (0 or 1)

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

1.2.2. Univariate Analysis:

After data preprocessing, we perform univariate analysis for the dataset variables. Each variable is univariately summarized with summary statistics, histograms, box plots and density plots to explore their distribution and characteristics. Such knowledge base of the distribution and variation in each attribute allows us to conduct more extensive analysis or hypothesis-evaluation.

- **Gender Distribution**

We examine the gender distribution of patients in the dataset. We bar plot gender composition in terms of frequency illustrated in *Figure 2*. Aggregated statistics such as counts do not adequately reflect gender distribution in the group, and this information is key to monitoring diabetes distribution by genders.

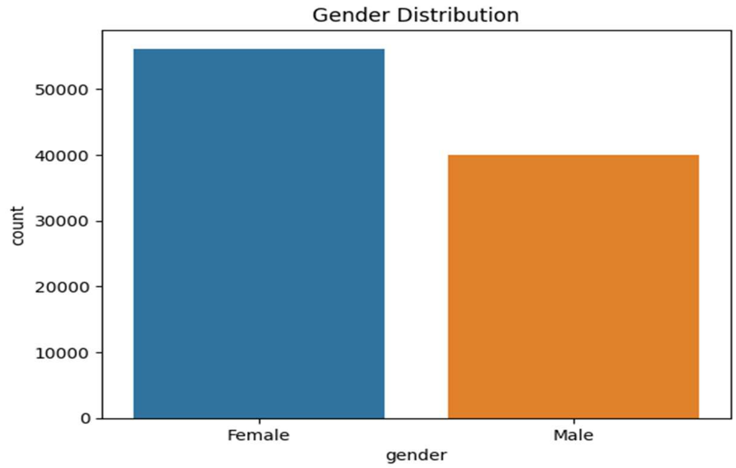

Figure 2: Gender distribution

- **BMI Distribution**

We examine patient BMI distribution demonstrated in *Figure 3*. To do this we examine central tendencies and variability in a BMI distribution through density plots. An advantage of mean, median, and percentiles is in summarizing numerically. Knowledge about general weight distribution in patients with diabetes is essential for estimating a risk of disease.

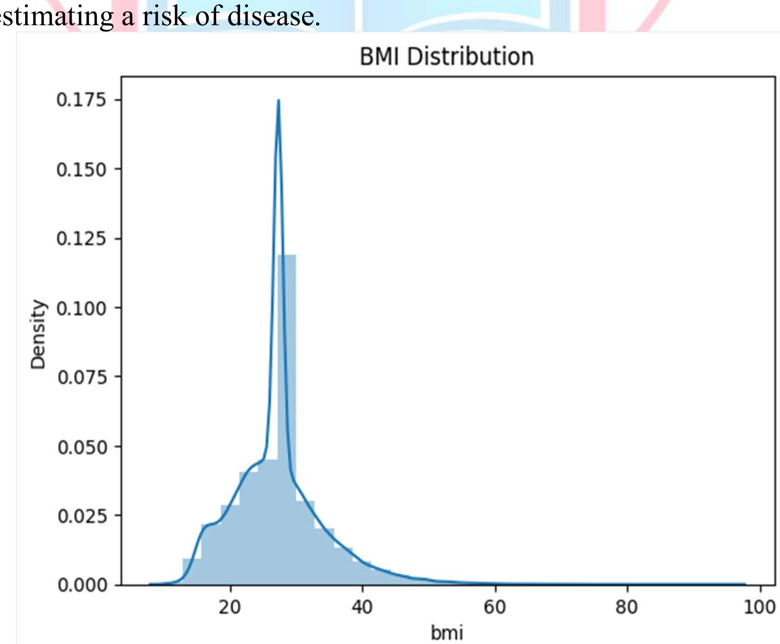

Figure 3: BMI distribution

- **Smoking History Distribution**

The distribution of smoking history categories is analyzed to assess the prevalence of smoking among patients. Bar charts are used to visualize the different smoking categories, while frequency counts and proportions provide quantitative insight. This analysis helps inform targeted treatment strategies by highlighting smoking patterns among diabetic patients and understanding their potential impact on disease progression and related complications you can see in *Figure 4*.

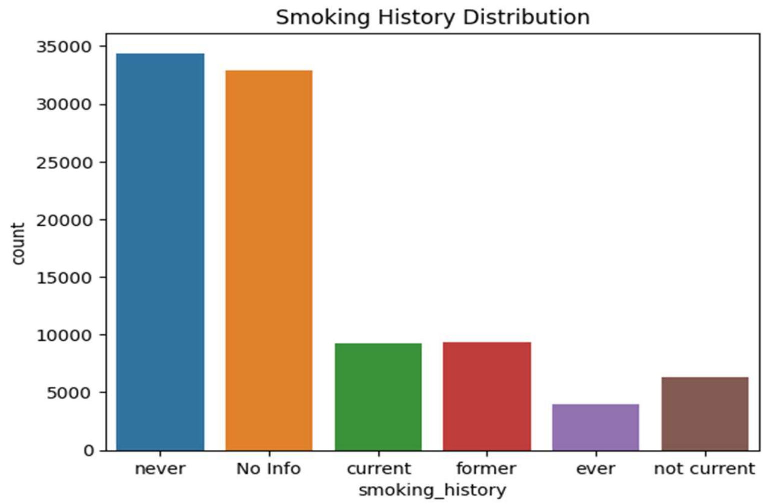

Figure 4: Smoking History Distribution

1.2.3. Bivariate Analysis

Once we have done the univariate analysis of individual variables, we now go further and do the bivariate analysis to investigate the relationship between pairs of variables. Bivariate is used to determine the relationships between variables and determine how strong and in which direction the relationship is. We plot and measure the associations between variables with scatter plots, correlation matrices, and contingency tables to discover the trends and relationships that can possibly influence the prediction of diabetes.

- **BMI vs. Diabetes Classification**

This paper investigates the correlation between BMI and diabetes detail given in *Figure* 5. Box plots depict the difference between diabetic and non-diabetic BMI. The difference in BMI is measured by t-tests or by Mann-Whitney U test. This demonstrates the influence of BMI on the risk of diabetes.

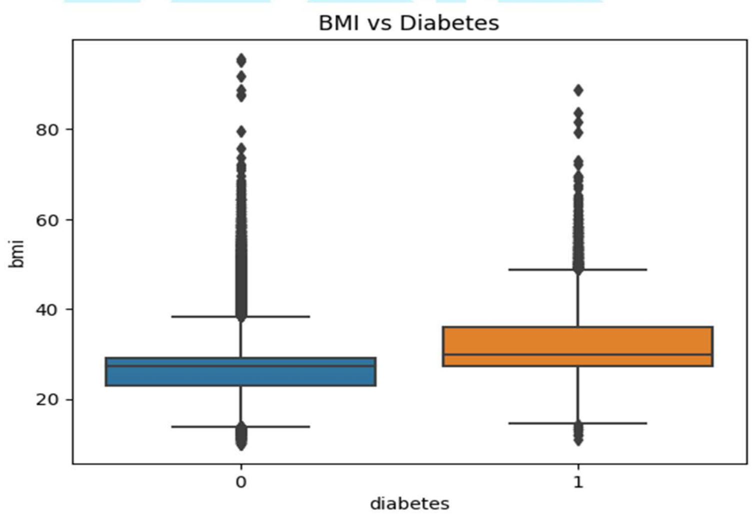

Figure 5: BMI vs Diabetes Classification

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

- **Blood Glucose Level vs. Diabetes Classification**

We check the level of glucose in the blood and diabetes classification. Box plots are used to plot diabetic and non-diabetic glucose levels. Mean glucose is compared by T- and Wilcoxon rank-sum tests. This discussion demonstrates the diabetes detection capability of blood glucose readings details are available in Figure 6.

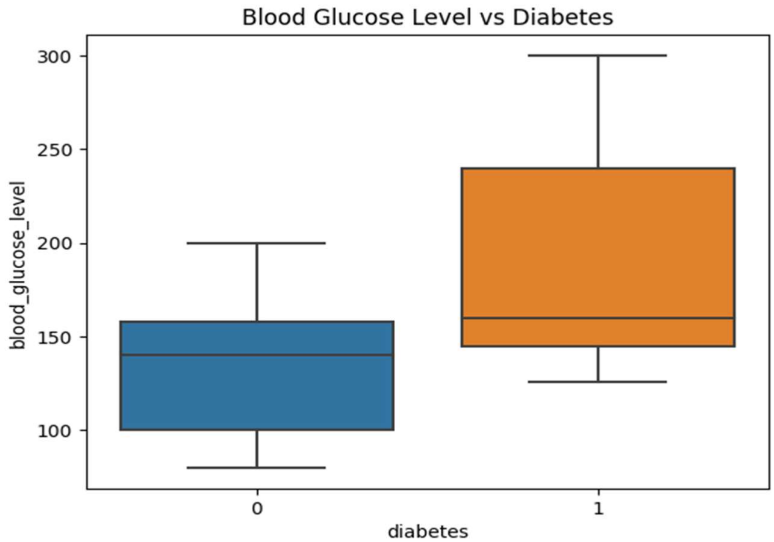

Figure 6: Blood Glucose Level vs Diabetes Classification

1.2.4. Multivariate Analysis

Multivariate analysis is applied in the last stage of data analysis to study the correlation between many variables. Multivariate analysis is used to identify complex data patterns and clusters to get a clearer insight into diabetes prediction.

- **Age vs. BMI Colored by Diabetes Classification**

This paper investigates age and BMI, with diabetes categorization as you can see in *Figure 7*. The relation between age and BMI and the color-coding by the presence of diabetes helps to plot the variables and their interaction. This gives a full image of age, BMI and diabetes classification, which show pertinent risk factors.

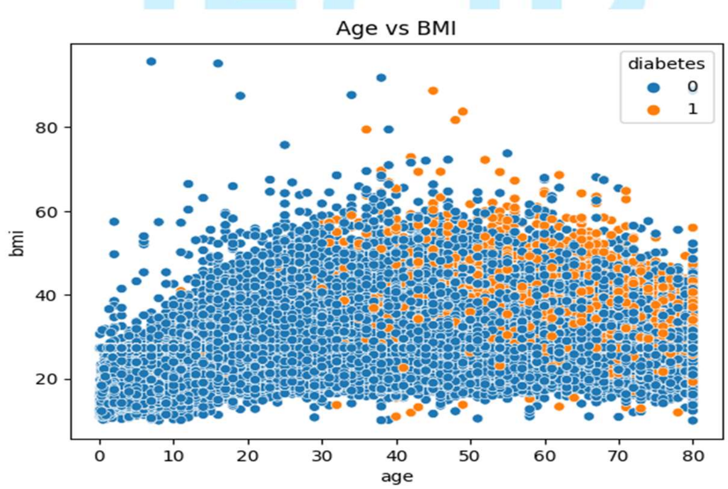

Figure 7: Age vs BMI by Diabetes

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

- **Age Distribution by Diabetes Status and Gender**

Age distribution by diabetes status and gender gives a subtle perception of the prevalence of diabetes among demographic groups. By visualizing non-diabetic age distributions and diabetes age distributions by gender, we will be able to see the differences and trends. It is a way that we can learn about diabetes epidemiology because this approach shows the differences between genders in terms of diabetes prevalence by the age group see *Figure 8* for more details.

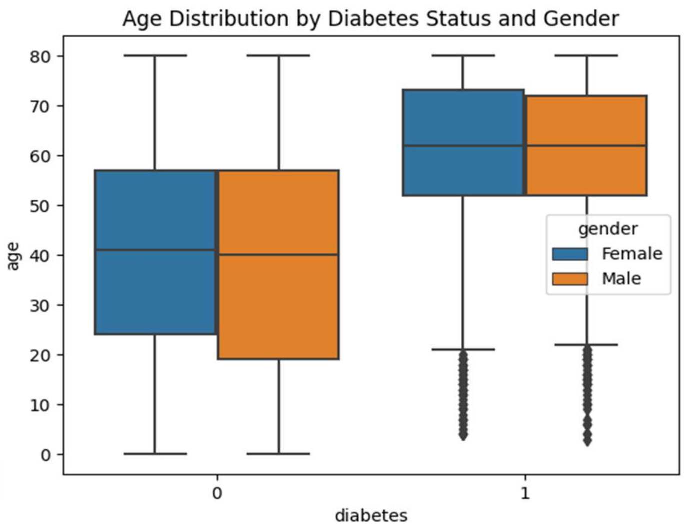

Figure 8: Age by Diabetes Status and Gender Correlation

The relationship between two or more variables is referred to as correlation. In addition to identifying the crucial features and cleansing the dataset prior to modelling, these steps also contribute to the model's efficiency as illustrated in *Figure 9.*

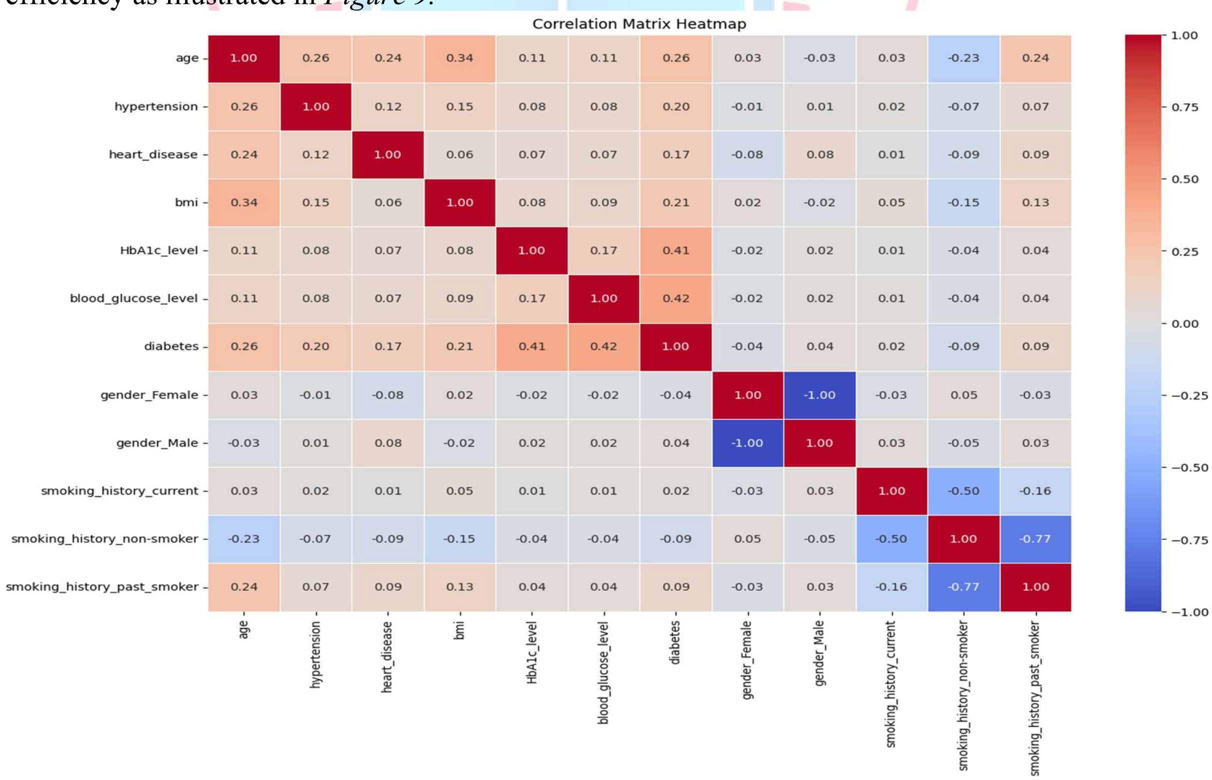

Figure 9: Correlation Representation

1.3. Model Building and Hyperparameter Tuning

We construct a pipeline that primitively preprocesses data and trains a model. A powerful classification method, Random Forest Classifier, is utilized by us to enhance the accuracy of predictions. The gridSearchCV is an exhaustive method of optimizing model hyperparameters. Cross-validation is used to choose the best model.

1.3.1. Preprocessing Steps

Before implementing these models, there are two pre-processing steps that follow us.

- ***Feature Engineering***

Feature engineering plays an important role in improving model performance; however, newly created variables must be meaningful and logically derived. In this dataset, additional features were generated based on BMI, insulin, and glucose values to better capture underlying patterns in the data.

- ***One Hot Encoding***

Since machine learning models require numerical inputs, all categorical variables in the dataset were transformed into numerical forms. This conversion was carried out using Label Encoding and One-Hot Encoding techniques, ensuring that categorical information could be effectively utilized by the model.

1.3.2. Data Spllitting and Training

The data is currently separated into training and testing datasets. Various models are trained and evaluated utilizing the training and assessment datasets, respectively. Additionally, cross-validation is executed on multiple models prior to making predictions on the testing data. The dataset was divided into two subsets: the test (20%) and the train (80%).

1.3.3. Hyperparameter Tunning

Hyperparameter tuning involves the repetitive exploration of a set of hyperparameters that are predefined and which seek the optimum model performance. A common method of this search is gridSearchCV. In GridSearchCV, cross-validation of the performance of the model is done through the learning process. Cross-validation consists of dividing the training data into numerous folds and training the model with one-fold and testing it with the remaining folds. Every hyperparameter configuration is repeated, and each fold performance (e.g., accuracy, F1-score) averaged across all folds is used to assess its performance.

These parameters are a result of the Hyperparameter tuning process as you can see in *Figure 10* (b), and they give us insight into the structure of the data and the complexity of the model that best captures that structure. The moderately constrained tree depth and the requirements for the number of samples at each node suggest a model that is complex enough to capture the important patterns in the data as shown in

Figure (a), but not so complex that it overfits to noise or outliers. This balance is crucial in creating a model that will generalize well to new data.

It is important to note that these parameters are optimal only within the defined parameter grid and for the specific dataset used in this study. If a different dataset is employed or the parameter grid is modified, the resulting optimal parameter values may vary accordingly.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

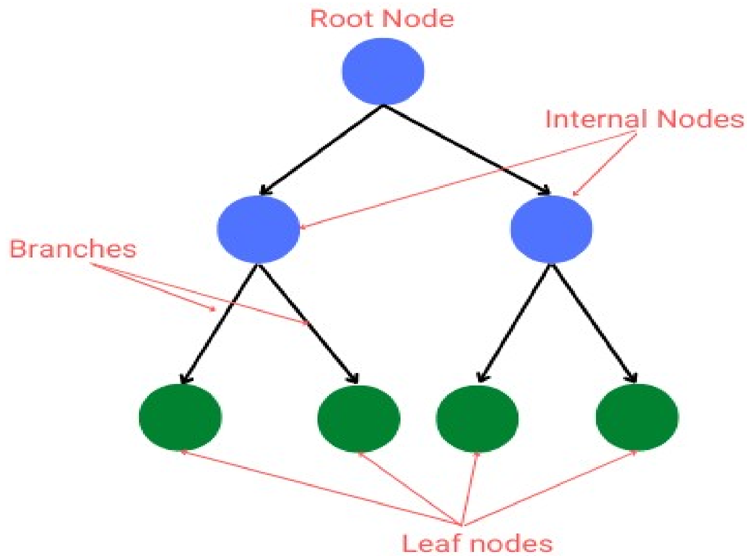

Figure 10: (a) Hyper parameter tuning

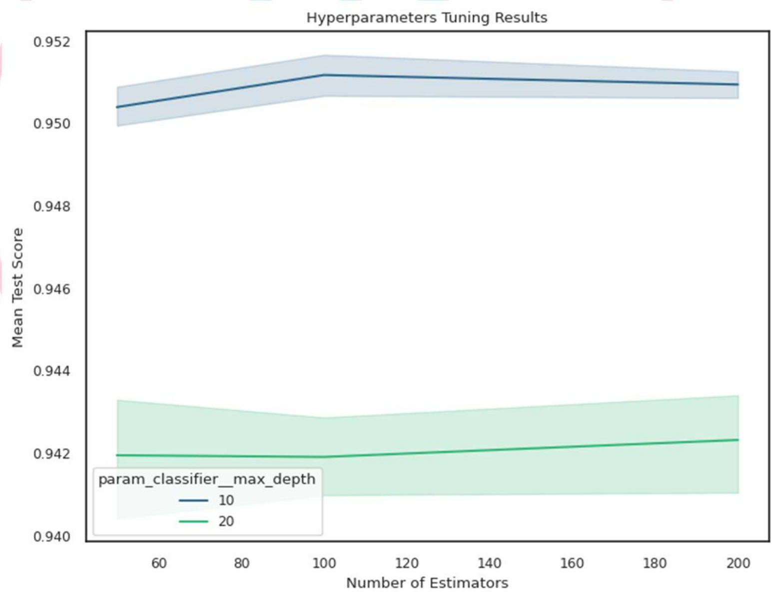

Figure 10: (b) Hyperparameters Tunning Results

1.4.Evaluation Matrix

The test set is used to evaluate the performance of the trained model on unseen data. A confusion matrix is employed to present the model's classification results in a clear and structured manner. It summarizes the outcomes in terms of true positives, true negatives, false positives, and false negatives. True Positives (TP) refer to instances where the model correctly predicts the presence of diabetes. True Negatives (TN) indicate cases where the model correctly predicts the absence of diabetes. False Positives (FP) occur when the model predicts diabetes even though it is not actually present, while False Negatives (FN) represent cases where the model fails to detect diabetes despite its presence. This breakdown provides a detailed understanding of the model's strengths and weaknesses in making accurate predictions.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

Figure 11: Evaluation Matrix

Results and Discussion

The results of the evaluation display diabetes prediction model performance. The model has an accuracy of 94.70 with the examples. In non-diabetic examples (class 0), it was very precise, recalls and F1-score were high, which suggests good identification. The accuracy of predicting diabetes examples (class 1) is less precise, recall, and F1-score than the accuracy of class 0. Although this is the case, the model can recall diabetic cases at 81 percent. What these results indicate are that the model is quite effective in identifying non-diabetics but requires more analysis and development to be effective in identifying diabetic ones. *Figure 12 (a), (b)* and *Figure 13* represent the general results and the performance of the model. In addition, the loss curve provides valuable insights into the model's learning process. The loss graph demonstrates a steady decline during training as shown in *Figure 12 (a)*, indicating effective optimization and convergence. A lower loss value signifies reduced errors in prediction, reinforcing the model's reliability. The alignment between accuracy improvement and loss reduction confirms the model's robustness in detecting diabetes-related patterns. However, minor fluctuations in the loss curve might suggest areas for further fine-tuning, such as adjusting hyperparameters or incorporating additional features for enhanced generalization.

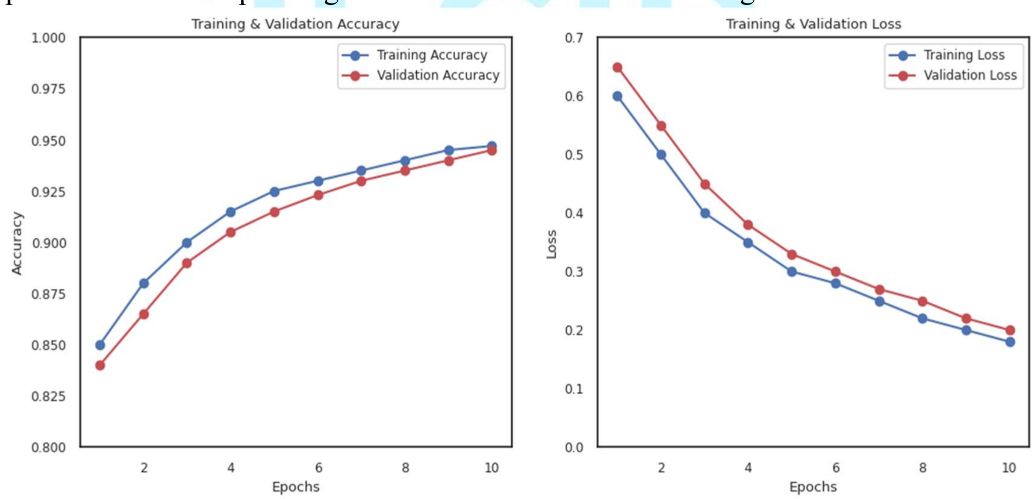

Figure 12: a) Model accuracy and loss

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

```
Model Accuracy:  0.9469988557162177
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     17525
           1       0.67      0.81      0.73      1701

    accuracy                           0.95     19226
   macro avg       0.82      0.88      0.85     19226
weighted avg       0.95      0.95      0.95     19226
```

Fig12: b) Model accuracy

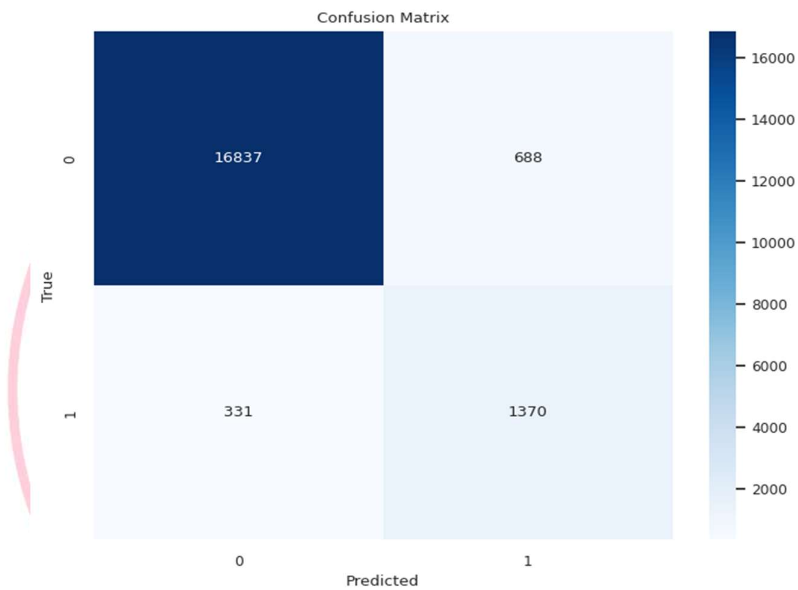

Figure 13: Confusion Matrix

The Random Forest classifier trained on the dataset achieved an overall accuracy of 95%, meaning that approximately 95% of the test samples were correctly classified. A closer analysis of the classification metrics for each class provides deeper insight into the model's performance.

- For Class 0 (non-diabetes), the model performs very well, achieving a precision of 0.98, which indicates that most predictions labeled as non-diabetic are correct. The recall of 0.96 shows that the model successfully identifies 96% of actual non-diabetes cases. Consequently, the F1 score of 0.97 reflects a strong balance between precision and recall for this class.
- For Class 1 (diabetes), the model shows moderate precision and strong recall. A precision of 0.65 means that 65% of the cases predicted as diabetic are truly diabetic. However, the recall of 0.80 indicates that the model correctly detects about 80% of actual diabetes cases, which is particularly important in healthcare applications. The resulting F1 score of 0.72 highlights this trade-off between precision and recall.

Overall, the weighted average F1 score of 0.94 aligns closely with the model's accuracy, confirming its robust performance. Importantly, the relatively high recall for the diabetes class is a positive outcome, as reducing false negatives; missed diabetes cases; is critical in medical decision-support systems.

CONCLUSION

In conclusion, this study aims to reach the objective of creating a reliable model of diabetes prediction by relying on a dataset of a well-known repository Kaggle. With an exceptionally high overall accuracy of approximately 94.70% achieved through meticulous preprocessing of the data, feature design and training on

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN: 3105-4811
Volume 3, Issue 2, 2025
https://jecir.com/

a Random Forest Classifier, the model can be trained to achieve significantly high accuracy. The effectiveness of this technique in terms of the prediction of diabetic cases is greatly reduced despite the fact it has shown good precision, recall and F1-score in terms of non-diabetic cases and it means that it is effective in identification. The possibility of the model to precisely classify the incidences of diabetes, which is less precise and with a low recall, suggests that there could be room for development. There is need to conduct further studies and refine the model to enhance the level of prediction of diabetic cases. This will eventually result in an effective evaluation and care of the risk of diabetes in the health care facility.

REFERENCES

[1] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Inform. Med. Unlocked*, vol. 10, pp. 100–107, 2018.

[2] M. M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*. Singapore: Springer, 2020, pp. 113–125.

[3] F. G. Woldemichael and S. Menaria, "Prediction of diabetes using data mining techniques," in *2018 2nd Int. Conf. Trends Electron. Inform. (ICOEI)*, May 2018, pp. 414–418.

[4] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 449–457, 2019.

[5] A. Aldallal and A. A. A. Al-Moosa, "Using data mining techniques to predict diabetes and heart diseases," in *2018 4th Int. Conf. Frontiers Signal Process. (ICFSP)*, Sep. 2018, pp. 150–154.

[6] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, "Detection and prediction of diabetes using data mining: a comprehensive review," *IEEE Access*, vol. 9, pp. 43 711–43 735, 2021.

[7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.

[8] P. Kumar, P. Kumar, A. Srivastava, V. D. A. Kumar, K. Vengatesan, and A. Singhal, "Comparative analysis of data mining techniques to predict heart disease for diabetic patients," in *Adv. Comput. Data Sci.*, Singapore: Springer, Apr. 2020, pp. 507–518.

[9] T. R. Mahesh et al., "Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease," *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[10] N. Oza and A. Bokhare, "Diabetes prediction using logistic regression and K-nearest neighbor," in *Congr. Intell. Syst.*, Singapore: Springer, 2022, pp. 407–418.

[11] K. S. Anil and R. Jain, "Data mining techniques in diabetes prediction and diagnosis: a review," in *2022 6th Int. Conf. Trends Electron. Inform. (ICOEI)*, Apr. 2022, pp. 1696–1701.