International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

An Advanced Machine Learning Approach for Enhanced Diabetes Prediction

Almonzer Salah Nooraldaim[1,2], Amal Elobaid Ahmed Abdalla[2], Amna Mirghani Seed[3]
[1]Department of Computer Science, Xi'an Jiaotong University, Xi'an, China
[2]School of Pharmacy, International University of Africa, Khartoum, Sudan
[3]Sudan University of Science and Technology, Department of Computer Science, Khartoum, Sudan

ABSTRACT: Diabetes is a chronic health condition affecting millions globally, causing severe complications and burdening healthcare systems. Current machine learning methods for diabetes prediction face challenges such as data imbalance, limited generalizability, and computational inefficiency. This study proposes a novel method that combines K-Nearest Neighbors (KNN), clustering techniques, Synthetic Minority Over- sampling Technique (SMOTE), and Random Forest for outcome classification to address these issues. The PIMA Indian Diabetes Dataset was used to evaluate the approach, achieving accuracy of 87.50%. However, the study has limitations, such as dependency on specific datasets and computational complexity. Future work will focus on validating the method across diverse datasets, optimizing computational efficiency, and developing real-time prediction capabilities.

KEYWORDS: Data Imbalance, Diabetes Prediction, Healthcare, Machine Learning, Random Forest, Synthetic Minority Over-sampling Technique.

1 INTRODUCTION

Diabetes, a chronic medical condition that causes severe damage to organs and systems[1],[2], is a global health crisis that affects millions of people around the world[3]. The increasing incidence of diabetes calls for improved management strategies and international efforts to curb its progression and mitigate its public health repercussions[4]. Diabetes is a complex condition characterized by three types: Type 1, Type 2, and gestational diabetes[5, 6]. Type 1 is an autoimmune disease that causes the body to fail to produce insulin, leading to high blood sugar levels[7]. Type 2 diabetes is more common and results from the body's inability to effectively use insulin[8]. Gestational diabetes, a temporary condition during pregnancy, increases the mother's risk of developing Type 2 diabetes later[9]. Understanding these differences is crucial for effective care and improved patient outcomes[10]. In recent years, diabetes has become a global health issue, driven by factors such as sedentary lifestyles, unhealthy diets, and an aging population[11],[12]. This rapid rise places immense pressure on healthcare systems and leads to substantial social and economic burdens worldwide[13]. Effective strategies for early detection and management are essential to mitigate complications such as cardiovascular disease, kidney failure, and neuropathy. Addressing these challenges requires a combination of traditional healthcare approaches and innovative technologies to improve prevention, diagnosis, and treatment outcomes[14],[15].

Machine learning (ML) is a crucial tool in medical research, particularly in the treatment of diabetes[16]. It can analyze large datasets, predict disease onset, and improve treatment effectiveness[17]. ML algorithms can uncover hidden patterns, leading to the development of personalized medicine and preventive care strategies, which can significantly improve patient outcomes and reduce disease incidence[18],[19],[20]. The PIMA Indian Diabetes Dataset is a widely used resource in diabetes research, focusing on a cohort of Pima Indian women in Arizona, USA. It includes diagnostic measurements such as glucose con- centration, BMI, age, and insulin levels, making it an excellent case study for machine learning techniques. However, the data set faces challenges such as data imbalance, potential biases, and generalizability. Careful data handling and analysis are needed to ensure accurate insights and apply them to other populations, enhancing the utility of the data set in the development of targeted interventions for the management of diabetes.

The aim of this paper is to design a novel method that enhances existing approaches by combining techniques such as K-Nearest Neighbors (KNN) for imputation, clustering, Synthetic Minority Over-sampling Technique (SMOTE) for

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

improved data balancing, and classifying the outcome. This integrated approach aims to address the challenges of data quality, imbalance, and accurate prediction, ultimately improving the effectiveness of diabetes management models. The organization of this paper is as follows. Section 2 discusses related work. Section 3 presents the proposed methods. Section 4 showcases the results. Section 5 provides the discussion. Finally, Section 6 concludes the paper.

2 RELATED WORK

Several studies have explored the application of machine learning techniques to diabetes prediction and management, highlighting the potential of these methods to improve health- care outcomes. By leveraging datasets such as the PIMA Indian Diabetes Dataset, re- searchers have tested various classifiers and feature selection approaches to enhance the accuracy and reliability of predictive models. These efforts underscore the growing importance of machine learning in addressing the global diabetes challenge. This study [21]compared various classifiers and feature selection techniques using the Indian Diabetes Dataset PIMA, focusing on random forest classifiers. After extensive pre-processing and hyperparameter optimization, the random forest classifiers achieved the highest accuracy at 79.80%. However, another study [22] developed an intelligent framework for predicting diabetes using machine learning techniques, leveraging big data analytics and the Pima Indian Diabetes Database. The framework achieved 83.00% accuracy with minimal error rates, demonstrating potential for healthcare professionals and researchers.

Additionally, this paper [23] proposed a machine learning method to improve diabetes diagnosis, using algorithms such as random forest, multilayer perception, logistic regression, and LSTM. The study found that multilayer perception outperformed other classifiers and introduced an IoT-based glucose monitoring model and achieved 80.84%. Furthermore, an- other study by Md. Maniruzzaman [24] compared traditional diabetes classification methods using Gaussian Process Classification (GPC) with radial basis kernels. The results showed that the GPC model with the radial basis kernel outperformed others, achieving an accuracy of 81.97%, with a sensitivity of 91.79%, and a specificity of 63.33% [24]. a separate study conducted an extensive study with three classifiers: random forest, multilayer perceptron, and logistic regression. Their study demonstrated the superior performance of the multilayer perceptron classifier, achieving an accuracy of 86.06% [25].

Consequently, Therefore, a study using logistic regression and decision tree models predicted type 2 diabetes in Indian women using the PIMA dataset. The model identified five main predictors: glucose, pregnancy, BMI, diabetes pedigree function, and age, and achieved an accuracy of 78.26%, highlighting its effectiveness in identifying high-risk individuals [26]. In turn, A research by Gnanadass [27] addressed the issue of missing data by imputing the mean for each column. Six distinct models were trained, with the XG- Boost model attaining the greatest accuracy rate of 77.54%. Hayashi and Yukita [28] proposed employing a rule extraction method, Re-RX, integrated with J48 graft and a sampling selection technique, to attain an accuracy of 83.83%.The summary of various studies highlighting machine learning approaches for diabetes prediction and their respective contributions is presented in Table 1.

3 MATERIALS AND METHODS

3.1 Dataset

The PIMA Indian Diabetes Dataset is a public health dataset derived from a larger study conducted on the Pima Indian population near Phoenix, Arizona. This dataset is crucial for diabetes research due to the high prevalence of diabetes within this community, which is significantly higher than the global average. It contains data from 768 female patients of Pima Indian heritage. The dataset includes diagnostic measurements and medical details specifically related to diabetes. The variables recorded encompass showed in Table 2.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Table 1. Summary of machine Learning methods applied in diabetes prediction

Author(s)	Method Used	Results
Roshi Saxena, et al.[21]	MLP, decision trees, K-NN, random forest	Random forest achieved the highest accuracy of 79.8%.
Raja Krishnamoorthi, et al. [22]	Decision trees, random forest, SVM	Achieved 83% accuracy with minimal error rates for healthcare professionals.
Quan Zou, et al. [23]	Decision trees, random forest, neural networks with PCA	Random forest with all features gave an accuracy of 80.84%.
Md. Maniruzzaman, et al. [24]	GPC with linear, polynomial, and RBF kernels	Radial basis kernel outperformed LDA, QDA, and Naive Bayes with 81.97% ac- curacy.
Umair Muneer Butt, et al. [25]	Random forest, MLP, logistic regression, LSTM	MLP and LSTM achieved accuracies of 86.08% and 87.26%.
Ram D. Joshi, et al. [26]	Logistic regression and decision trees	Logistic regression predicted with 78.26% accuracy. Key predictors: glucose, BMI, age.
Gnanadass [27]	Naive Bayes, linear regression, random forest, AdaBoost gradient boosting machine, extreme gradient boosting	Achieve an accuracy of 78.00%.
Hayashi and Yukita [28]	J48 graft, rule extraction	J48 graft predict an accuracy of 83.83%.

3.2 Data preparation and preprocessing

We embarked on a two-pronged approach to enhance the data quality of the PIMA Indian Diabetes Dataset. Initially, we conducted a comprehensive cleaning process in which all corrupted or irrelevant records were removed. This step was critical in ensuring the standardization of the data, which is paramount for subsequent analyzes. Subsequently, we addressed the issue of missing data, which was prevalent across several key variables

Table 2. Features and descriptions for diabetes prediction

Feature	Description
Pregnancies	Number of times pregnant.
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test.
Blood pressure	Diastolic blood pressure (mm Hg).
Skin thickness	Triceps skin fold thickness (mm).
Insulin	2-Hour serum insulin (mu U/ml).
BMI	Body mass index (weight in kg/(height in m)^2).
Diabetes pedigree function	A function that scores the likelihood of diabetes based on family history.
Age	Age (in years).

Table 3. Dataset attributes, missing Values, and summary statistics

Attribute	Missing Values	Median	Mean
Pregnancies	0	3.00	3.85
Glucose	5	117.00	120.89
Blood Pressure	35	72.00	69.11
Skin Thickness	227	23.00	20.54
Insulin	374	30.50	79.80
BMI	11	32.00	31.99
Diabetes Pedigree Function	0	0.37	0.47
Age	0	29.00	33.24

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Within the dataset as shown in Table 3. To address this, we implemented the nearest neighbor (KNN) imputation method. The KNN imputation technique involves identifying the k-nearest neighbors to a record with missing values, based on the Euclidean distance, and imputing values by averaging the attributes of these neighbors see Figure 1. This method is particularly advantageous, as it maintains the inherent relationships between the data points, which is crucial for developing an accurate and reliable diabetes prediction model. This rigorous preprocessing strategy not only prepared the data set for effective analysis but also ensured the reliability of our predictive results, thus solidifying the foundation for applying complex machine learning algorithms to predict the onset of diabetes in the Pima Indian population.

3.3 Data imbalance

In addressing the challenge of data imbalance within the PIMA Indian Diabetes Dataset, we implemented a sophisticated approach to ensure a balanced distribution of data points across outcomes. This was crucial for avoiding bias in our predictive modeling, particularly given the propensity for imbalanced datasets to skew the performance of machine learning algorithms toward the majority class.

3.4 Dataset clustering

First up, we did K-means clustering on the dataset to partition it into clusters based on the proximity of data points. This had enabled us to group the data based on similar characteristics, which also helped us in dealing with imbalance within each group individually.

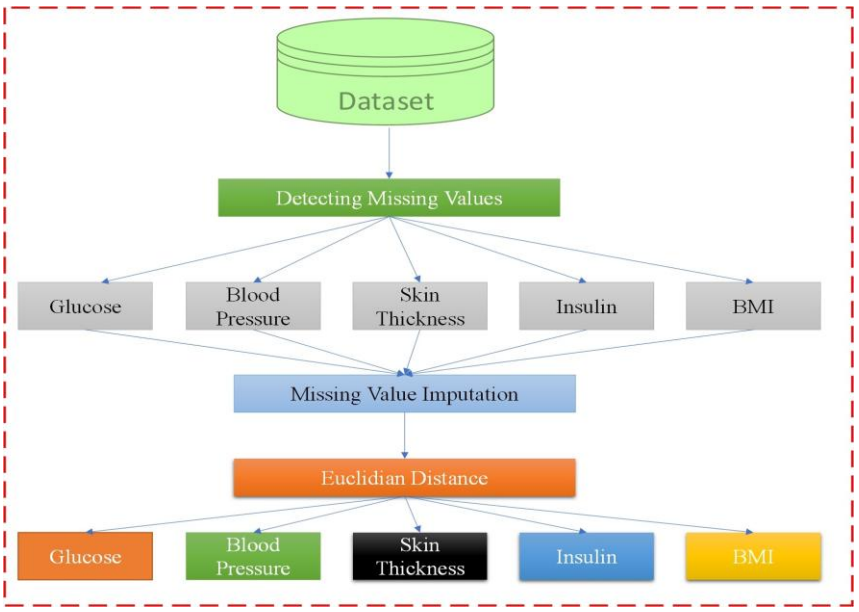

Fig. 1. Dataset using ecludian distance.

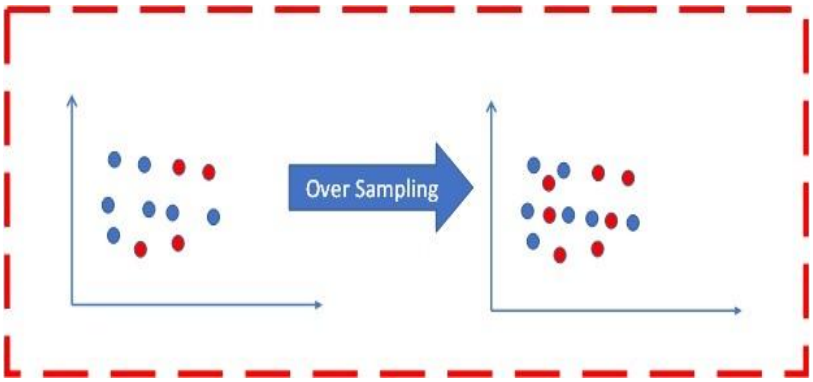

Fig. 2. Dataset using ecludian distance.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

3.5 Synthetic minority over-sampling technique(SMOTE)

After clustering, we applied the Synthetic Minority Over-sampling Technique (SMOTE) within each cluster. SMOTE is an advanced over-sampling method where synthetic samples are generated for the minority class, rather than creating copies of existing samples. This technique helps in providing a balanced representation by augmenting the minority class in each cluster with new, synthetic examples, derived from the feature space of the minority class members see figure 2. Once the clusters were individually balanced through SMOTE, we merged them back into a single, cohesive dataset, now with a significantly improved balance between the classes as show in figure 3. The final step in our data preparation process was to split this newly balanced dataset into training and testing sets. We allocated 70% of the dataset for training our machine learning models, ensuring they learn from a representative and comprehensive set of data points. The remaining 30% was set aside for testing, allowing us to evaluate the effectiveness of our models on previously unseen data, thereby providing a robust measure of model performance in real-world scenarios. This methodical approach to managing data imbalance not only enhanced the quality of our training and testing datasets but also ensured that the predictive insights derived from our study are both reliable and generalizable across various scenarios involving the prediction of diabetes onset.

We implemented the RandomForest classifier due to its robustness against overfitting and its capability to handle both numerical and categorical data efficiently. Each decision tree in the RandomForest ensemble operates on a random subset of data features and instances, resulting in varied learning perspectives that collectively enhance the model's generalization capabilities. This diversity makes the RandomForest particularly adept at dealing with the imbalanced data that had been meticulously balanced in the preprocessing stages of our research.

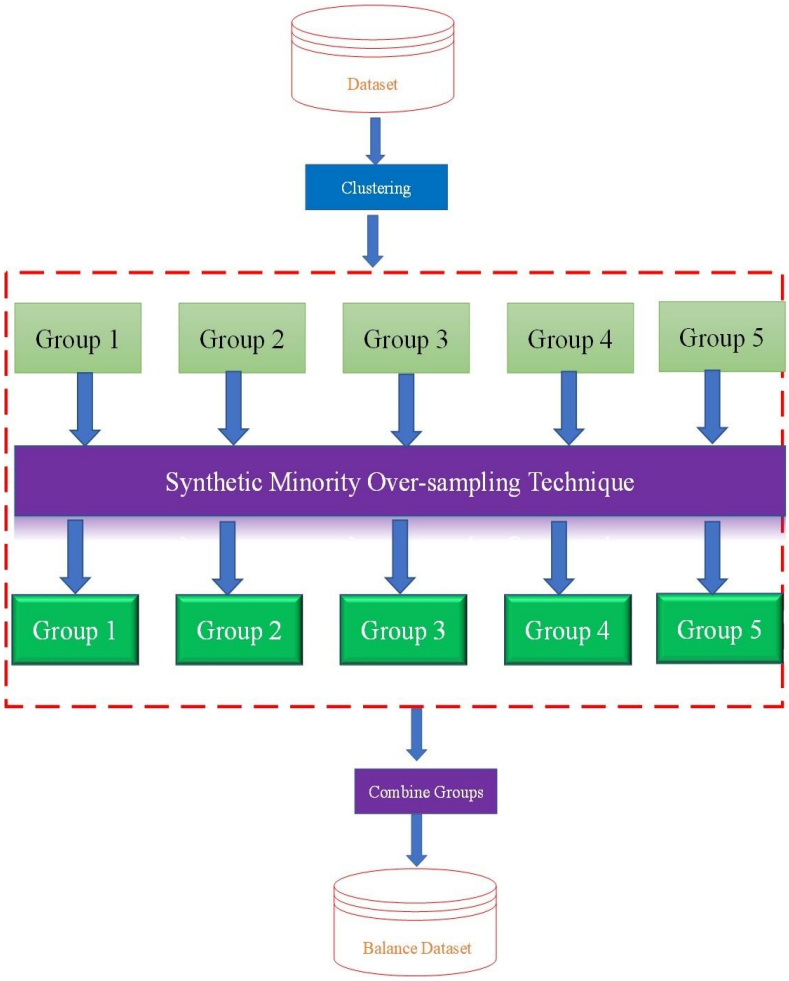

Fig. 3. The proposed method framework.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

4 RESULTS

4.1 Result of data preprocessing

The data preprocessing process focused on handling missing values for several key features without altering their natural distribution patterns that present in figure 4. For the Pregnancies, Glucose, and BloodPressure features, the data's overall shape and trends remained stable after filling in the missing values. Pregnancies showed a high frequency of zero values, which stayed consistent, while Glucose and BloodPressure maintained their normal distributions around 100-120 and 70-80, respectively. This consistency suggests that the imputation method preserved the original data structure for these features.

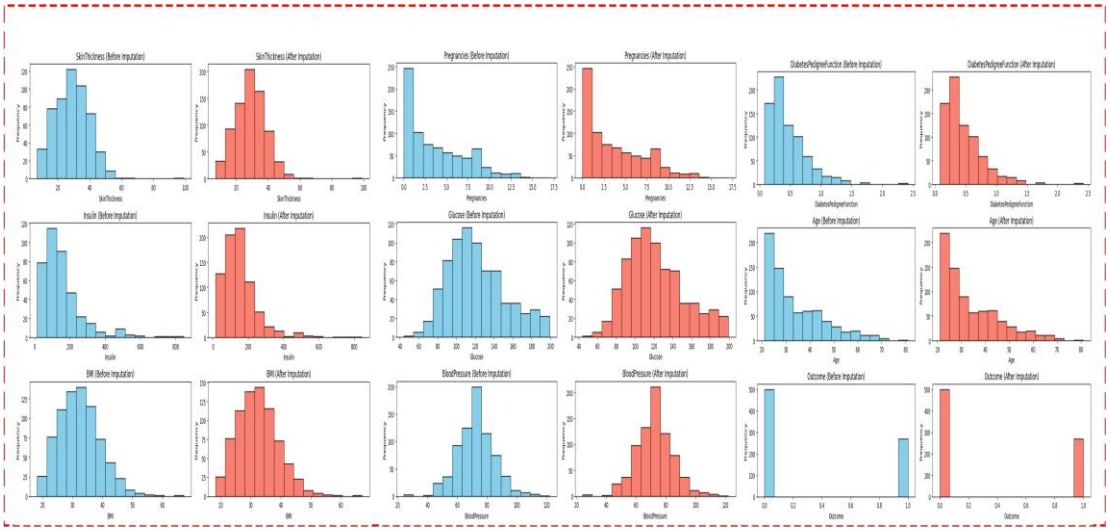

Fig. 4. The distribution of the data before and after imputation.

In the SkinThickness, Insulin, and BMI group, similar stability was observed. SkinThickness had a slightly skewed distribution towards lower values, which stayed unchanged after filling in gaps. Insulin showed a strong right skew with many low values and fewer high values; this skewness was preserved post-imputation, maintaining the original variability. BMI retained its bell-shaped distribution around a typical value of 30, showing that filling in missing values did not affect its central tendency or spread. Finally, the DiabetesPedigreeFunction, Age, and Outcome group exhibited the same patterns before and after imputation. DiabetesPedigreeFunction kept its right-skewed shape, Age continued to skew towards younger ages, and Outcome maintained its imbalance, with a majority of values at zero. The uniformity of this across features indicate that the preprocessing method effectively imputed missing values without perturbing the natural patterns of data. This stability of distributions post-imputation implies that the dataset is reliable for fur- ther analysis or modeling, as the statistical properties of the original data are effectively retained.

4.2 Model Performance Evaluation

We evaluated the performance of Random Forest classifier in classifying outcomes. Figure 5 shows the resulting predictions given a set of true and false values. The model classified 113 cases for the negative class (0) and 153 cases for the positive class (1). There were, however, 18 false positives (cases falsely predicted as positive) and 20 false negatives (cases falsely predicted as negative). This True or False prediction table provides information about the precision and recall of the model while providing us a sense of its ability to separate the two classes (although it could be better at being sensitive and specific!). The Random Forest classifier makes its final prediction based on the majority vote across multiple decision trees. For a given input X, the final prediction $y\hat{}$ of the Random Forest is determined by:

$$y\hat{} = \text{mode}\,(T_1(X), T_2(X), \ldots, T_n(X))$$

where:

- $T_i(X)$ is the prediction of the i-th decision tree for the input X,

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

- n is the total number of trees in the forest, and
- mode represents the most frequent class among the predictions, ensuring that the Random Forest model aggregates the individual tree predictions to produce a robust final decision.

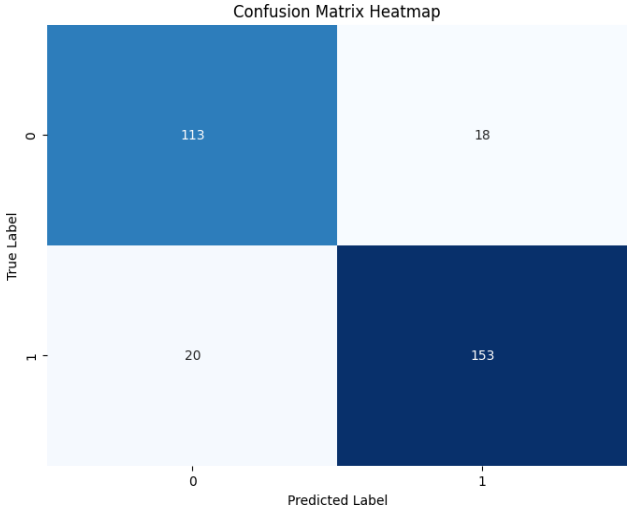

Fig. 5. Confusion matrix.

Random Forest trains each decision tree using a random subset of features and samples, reducing overfitted and improving the model generalization to new data. By using this ensemble approach the model is able to control for bias variance and be a good classifier across different scenarios, which you can see from the classification results.

4.3 Feature importance analysis

Feature importance analysis, as shown in Figure 6, allows us to see which of the features were most influential for the model decision process [41]. The strongest of these predictors was glucose, predicting that the value of the target variable would be relatively high. Additionally, BMI, Insulin and Age were also among the most important features in accordance with medical knowledge of their effects on health predicted outcomes. Features, such as SkinThickness, DiabetesPedigreeFunction, and BloodPressure had medium importance whereas Pregnancies were the least important features in the model. This ranking of feature importance helps guide further analysis, emphasizing the most relevant predictors to potentially enhance model performance in predicting health outcomes. The Random Forest model calculates feature importance based on the reduction in impurity (such as Gini impurity or Entropy) when a feature is used for splitting at a node. For a given feature X_j, the importance score $I(X_j)$ is computed as:

$$I(X_j) = \sum_{t=1}^{T} \Delta i_t(X_j)$$

where:

- T is the total number of trees in the Random Forest,
- $\Delta i_t(X_j)$ represents the reduction in impurity for feature X_j at node t.

This importance score reflects how much each feature contributes to decreasing impurity across all trees, guiding the model's split decisions. Features with higher $I(X_j)$ values are considered more important, as they lead to purer splits and contribute more to the predictive power of the model.

4.4 Comparison with other imputation methods

When K-Nearest Neighbors (KNN) was used for imputation, it provided the best results in terms of mean accuracy, reaching 87.50% as present in Table 4. This indicates that

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

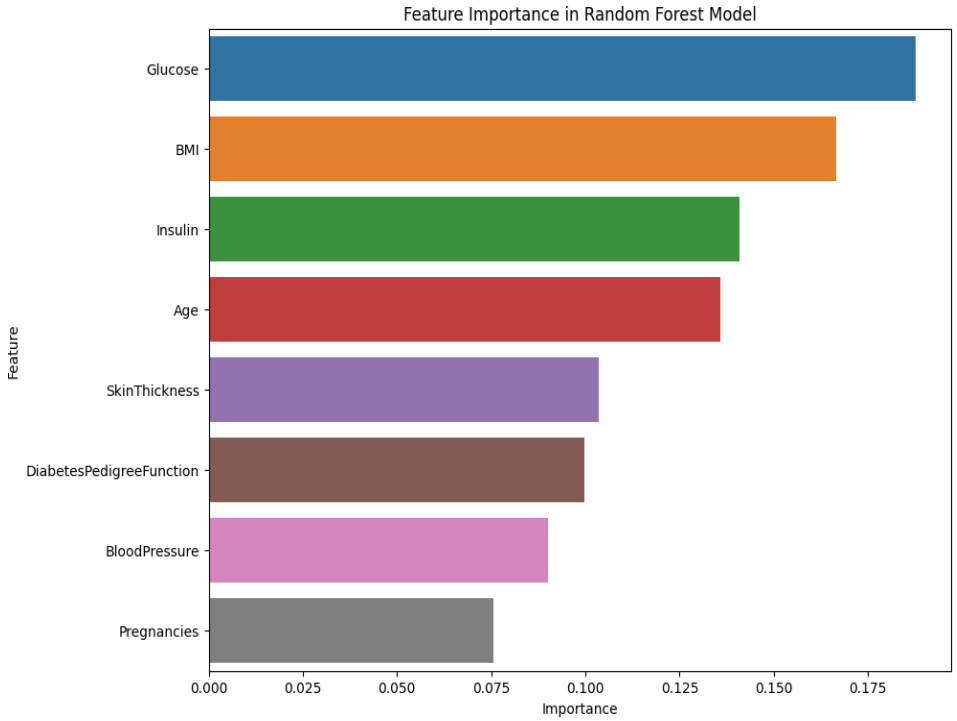

Fig. 6. Feature importance ranking.

KNN may be more effective in capturing underlying patterns in the data compared to other methods. Although it showed higher performance, its consistency was lower than Random Forest (RF), which had a mean accuracy of 83.37% and more stable results. In comparison, Support Vector Classifier (SVC) performed the worst, with a mean accuracy of 77.81%. Overall, KNN offers the best accuracy but may require further optimization to match the stability of methods like RF.

Table 4. Model Performance Comparison

Method	Mean	Median	Most Consistent	KNN
SVC	77.810	76.05	79.80	79.60
Neural Network	83.55	80.09	83.65	84.10
The proposed method	83.37	82.52	84.93	**87.50**

4.5 COMPARISON WITH OTHER CLASSIFICATION METHODS

The results in Table 5 highlight that the proposed method achieved the highest accuracy at 87.50%, outperforming various established models from previous studies. While Umair Muneer Butt et al.'s LSTM-based approach reached 87.26%, and Yoichi Hayashi et al.'s Recursive-Rule Extraction method attained 83.83%, the proposed approach showed a slight advantage. Several studies, including those by Roshi Saxena et al., Quan Zou et al., and Raja Krishnamoorthi et al., utilized Random Forest models with accuracies ranging from 79.8% to 83%. Additionally, Md. Maniruzzaman et al.'s Gaussian Process Classification reached an accuracy of 81.97%. These results underscore the effectiveness of the proposed method, which outperformed both traditional and specialized models in terms of accuracy, suggesting its potential for improved predictive performance across similar datasets.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Table 5. Accuracy comparison of different methods

Author	Accuracy (%)
Gnanadass [27]	78.00
RRam D. Joshi, et al. [26]	78.26
Roshi Saxena, et al.[21]	79.80
Quan Zou, et al. [23]	80.84
Md. Maniruzzaman, et al. [24]	81.97
Raja Krishnamoorthi, et al. [22]	83.00
Hayashi and Yukita [28]	83.83
Umair Muneer Butt, et al. [25]	87.26
Proposed Method	**87.50**

5 DISCUSSION

The proposed method, achieving an accuracy of 87.50%, demonstrates superior performance compared to most existing studies in diabetes prediction. Gnanadass [27] reported an accuracy of 78.00%, trailing the proposed method by 9.50%, while RRam D. Joshi et al.
[26] achieved 78.26%, with a difference of 9.24%. Similarly, Roshi Saxena et al. [21] reached 79.80%, which is 7.70% lower, and Quan Zou et al. [23] achieved 80.84%, showing a 6.66% gap. Md. Maniruzzaman et al. [24] performed better with 81.97%, but it still falls short by 5.53%. Raja Krishnamoorthi et al. [22] and Hayashi and Yukita [28] achieved accuracies of 83.00% and 83.83%, respectively, trailing by 4.50% and 3.67%. Umair Muneer Butt et al. [25], with an accuracy of 87.26%, closely approaches the performance of the proposed method but remains 0.24% lower. These results clearly establish the effectiveness of the proposed method as a highly accurate approach for diabetes prediction.
Feature importance analysis further supports the strength of the proposed method. As shown in Figure 6, Glucose emerged as the most influential feature, indicating its strong association with the target variable, which aligns well with clinical knowledge about glucose's role in health outcomes. Other features such as BMI, Insulin, and Age also ranked highly, reinforcing their significance in predicting health outcomes. Features like SkinThickness, DiabetesPedigreeFunction, and BloodPressure showed moderate importance, while Pregnancies had the least impact on the model's predictions. These findings indicate that the proposed method effectively identifies and leverages the most critical health-related predictors, supporting its superior performance. The ability to pinpoint such key features allows the model to focus on variables with the highest predictive power, enhancing its accuracy while potentially reducing computational complexity. This ranking aligns with medical insights, where glucose, BMI, and insulin are commonly associated with conditions like diabetes, underscoring the model's relevance and accuracy in health data contexts.
The proposed method demonstrates competitive performance when compared to other machine learning approaches as presented in Table 4, particularly in terms of consistency and overall accuracy. The Support Vector Classifier (SVC) achieved a mean accuracy of 77.81%, with a median of 76.05%, and its most consistent result was 79.80%, slightly improving to 79.60% under KNN imputation. However, these results fall significantly behind the proposed method. Neural Networks performed better, with a mean accuracy of 83.55% and a median of 80.09%. Its most consistent result reached 83.65%, and its accuracy increased to 84.10% under KNN imputation. While Neural Networks offer strong results, they still trail the proposed method in overall consistency and peak accuracy. The proposed method, with a mean accuracy of 83.37% and a median of 82.52%, achieved the highest consistent result at 84.93% and reached a maximum accuracy of 87.50%. This superior performance underscores the robustness of the proposed approach, which effectively combines feature selection and imputation techniques to outperform both SVC and Neural Networks, making it a highly reliable choice for diabetes prediction tasks.
Despite the high performance of our proposed method, we note its limitations. Firstly, its reliance on particular datasets

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

and population characteristics may restrict the generalizability to other datasets with diverse population traits. Moreover, while SMOTE was used to reduce data imbalance, there is a risk that this synthetic data may embed some bias that makes the model less robust. Also, the pinning could be found as infeasible in case of a larger dataset or more significant preprocessing process when more memory and CPU would be needed to run the proposed method. Improving these limitations would improve scalability and generalizability of the method to different contexts. Therefore, further research should validate the proposed method using more diverse datasets, implement advanced imputation-balancing techniques to avoid biases embedded in synthetic data, enhance levels of computational efficiency for deploying real-world healthcare systems seamlessly, and develop its future work by making it adaptive to predict in real-time scenarios as well as integrate automatic recommendations from clinical decision-support tools.

6 CONCLUSION

This study introduced a novel method that achieved 87.50% accuracy, surpassing existing models like LSTM and Recursive-Rule Extraction in health-related classification tasks. By focusing on key predictors, such as glucose, BMI, and insulin, the model demonstrated both accuracy and efficiency. Comparisons with other models confirmed its adaptability and strength in complex health data scenarios. While effective, future improvements could enhance interpretability and validate performance across diverse datasets, broadening the method's impact in predictive health analytics and clinical decision-making.

REFERENCES

1. National Diabetes Data Group (US), et al., Diabetes in America. No. 95, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 1995.
2. N. G. Forouhi and N. J. Wareham, Epidemiology of diabetes, Medicine, vol. 38, no. 11, pp. 602–606, 2010.
3. R. Bilous, R. Donnelly, and I. Idris, Handbook of Diabetes. John Wiley & Sons, 2021.
4. U. Alam, O. Asghar, S. Azmi, and R. A. Malik, General aspects of diabetes mellitus, Handbook of Clinical Neurology, vol. 126, pp. 211–222, 2014.
5. R. A. DeFronzo et al., Type 2 diabetes mellitus, Nature Reviews Disease Primers, vol. 1, no. 1, pp. 1–22, 2015.
6. E. Ginter and V. Simko, Type 2 diabetes mellitus, pandemic in 21st century, Diabetes: An Old Disease, a New Insight, pp. 42–50, 2013.
7. C. D. Deshmukh, A. Jain, and B. Nahata, Diabetes mellitus: A review, Int. J. Pure Appl. Biosci, vol. 3, no. 3, pp. 224–230, 2015.
8. S. C. Smith Jr, Multiple risk factors for cardiovascular disease and diabetes mellitus, The American Journal of Medicine, vol. 120, no. 3, pp. S3–S11, 2007.
9. J. H. Medalie, C. M. Papier, U. Goldbourt, and J. B. Herman, Major factors in the development of diabetes mellitus in 10,000 men, Archives of Internal Medicine, vol. 135, no. 6, pp. 811–817, 1975.
10. J. M. Olefsky, Prospects for research in diabetes mellitus, Jama, vol. 285, no. 5, pp. 628–632, 2001.
11. G. J. Shi et al., Involvement of growth factors in diabetes mellitus and its complications: A general review, Biomedicine & Pharmacotherapy, vol. 101, pp. 510–527, 2018.
12. Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Study Research Group, Long-term effect of diabetes and its treatment on cognitive function, New England Journal of Medicine, vol. 356, no. 18, pp. 1842–1852, 2007.
13. Diabetes Control and Complications Trial Research Group, The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus, New England Journal of Medicine, vol. 329, no. 14, pp. 977–986, 1993.
14. Howard, J. H. Arnsten, and M. N. Gourevitch, Effect of alcohol consumption on diabetes mellitus: A systematic review, Annals of Internal Medicine, vol. 140, no. 3, pp. 211–219, 2004.
15. S. A. Mazzuca et al., The diabetes education study: A controlled trial of the effects of diabetes patient education, Diabetes Care, vol. 9, no. 1, pp. 1–10, 1986.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 12 December 2024
DOI: 10.47191/ijcsrr/V7-i12-15, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

16. B. Mahesh, Machine learning algorithms-a review, International Journal of Science and Research (IJSR), vol. 9, no. 1, pp. 381–386, 2020.
17. Z. H. Zhou, Machine learning, Springer Nature, 2021.
18. M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science, vol. 349, no. 6245, pp. 255–260, 2015.
19. Singh, N. Thakur, and A. Sharma, A review of supervised machine learning algorithms, in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1310–1315, IEEE, Mar. 2016.
20. S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, Supervised machine learning: A review of classification techniques, Emerging Artificial Intelligence Applications in Computer Engineering, vol. 160, no. 1, pp. 3–24, 2007.
21. R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, A novel approach for feature selection and classification of diabetes mellitus: machine learning methods, Computational Intelligence and Neuro- science, vol. 2022, no. 1, pp. 3820360, 2022.
22. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, Predicting diabetes mellitus with machine learning techniques, Frontiers in Genetics, vol. 9, pp. 515, 2018.
23. R. Krishnamoorthi et al., [Retracted] A novel diabetes healthcare disease prediction framework using machine learning techniques, Journal of Healthcare Engineering, vol. 2022, no. 1, pp. 1684017, 2022.
24. M. Maniruzzaman et al., Accurate diabetes risk stratification using machine learning: role of missing value and outliers, Journal of Medical Systems, vol. 42, pp. 1–17, 2018.
25. U. M. Butt et al., Machine learning based diabetes classification and prediction for healthcare appli- cations, Journal of Healthcare Engineering, vol. 2021, no. 1, pp. 9930985, 2021.
26. R. D. Joshi and C. K. Dhakal, Predicting type 2 diabetes using logistic regression and machine learning approaches, International Journal of Environmental Research and Public Health, vol. 18, no. 14, pp. 7346, 2021.
27. Gnanadass, Prediction of gestational diabetes by machine learning algorithms, IEEE Potentials, vol. 39, no. 6, pp. 32–37, 2020.
28. Y. Hayashi and S. Yukita, Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset, Informatics in Medicine Unlocked, vol. 2, pp. 92–104, 2016.

Cite this Article: Almonzer Salah Nooraldaim, Amal Elobaid Ahmed Abdalla, Amna Mirghani Seed (2024). An Advanced Machine Learning Approach for Enhanced Diabetes Prediction. International Journal of Current Science Research and Review, 7(12), 8779-8789, DOI: https://doi.org/10.47191/ijcsrr/V7-i12-15