

Machine Learning for Diabetes Prediction Using Random Forest: A Comprehensive Analysis with Class Balancing Techniques

Hashim Ali¹, Asif Rahim², Aftab Ahmed³, Umer Tanveer⁴, Wali Samad Khan⁵,
Shiasta Bibi⁶

¹Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

²School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, 541004, China

³Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

⁴Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

⁵Department of Biotechnology, Abdul Wali Khan University, Mardan, Pakistan.

⁶Roots Millenium School Mardan, Pakistan.

DOI: <https://doi.org/10.5281/zenodo.18030488>

Keywords:

Diabetes prediction, feature importance, machine learning, medical diagnosis, Random Forest, classification, SMOTE, SMOTE-ENN, class imbalance

Article History

Received on 14 Oct, 2025

Accepted on 11 Nov 2025

Published on 13 Nov 2025

Copyright @Author

Corresponding Author:

Abstract

Diabetes is a major global prevalence. Health problems that affect millions of people round the world and placing significant burdens on health systems. Primitive diabetes risk detection and prediction can be significantly improved. Patient outcomes through timely intervention and management Strategies. This article provides a comprehensive analysis Predicting diabetes using machine learning techniques. Special Focus on Random Forest Algorithm and Class Balance Methods. We have a strong pre-processing pipeline that handles missing values that encodes variable category and removing irrelevant identifiers from the dataset. Addressing the imbalance class, there are two types of synthetic techniques: synthetic Minority oversampling Techniques (SMOTE) and SMOTE with edited nearest Neighbours (SMOTE-ENN). Our Forest classifier achieves notable performance metrics, with original model 99.00% accuracy, accuracy, remember and F1-score. Comparative analysis shows that when SMOTE (98.50% accuracy) and SMOTE-ENN (95.00% accuracy) reduce overall performance metrics; they provide more balanced representation of class. The analysis of the characteristics shows that HbA1c, BMI, and age are the most important predictors diabetes in all models. The learning curve and the chaos the matrix demonstrates the strong classification capabilities of the models and clarifies the compromises between class balance and general equilibrium execution. This research has led to the expansion of the body understanding the applications of machine learning by healthcare a detailed methodology for the diagnosis of diabetes it can also be applied to clinical decision support systems. Improving early diagnosis and patient management strategies.

Index Terms diabetes prediction, feature importance, machine learning, medical diagnosis, Random Forest, classification, SMOTE, SMOTE-ENN, class imbalance

I. INTRODUCTION

Diabetes is one of the most common diseases worldwide, affecting about 537 million people as of 2021, adults around the world, and forecasts show that This number is expected to rise to 783 million by 2045.[1] It's metabolic disorder, which is characterized by elevated blood glucose, has a condition these include major health problems, including an increased risk of heart and blood pressure disease, nephropathy, neuropathy, and retinopathy. The economic burden of diabetes is huge. The cost of diabetes is estimated at \$966 billion [2]. Early detection and prognosis of diabetes risk Strategies to Reduce the Effects of the Disease Timely interventions that can significantly improve the patient's health Outcomes and Reduced Health Costs. Traditional diagnosis Methods rely widely on clinical indicators such as fasting Plasma Glucose, Oral Glucose Tolerance Test, and Glycated Hemoglobin (HbA1c) levels. But these methods are often these: Diabetes is only detected when there is significant progress. The need for an active forecast [3]. Machine learning techniques have been shown to be remarkable. Potential in medical applications, especially in diseases prediction and risk classification. These computational methods can recognize complex patterns and relationships in medical matters. Information that may not be immediately disclosed through the traditional statistical analysis. In the midst of a variety of machine learning random forest has attracted a lot of attention due to its robustness, interpretability, and ability to handle high-dimensional behavior information that needs to be processed at least in advance [4].

This article provides a comprehensive analysis of diabetes. The prediction was made using a random algorithm through a systematic method that includes the pre-processing of data, model training, evaluation, and the importance of features analysis. We also address the serious challenge of class imbalance Through Two Advanced Technologies: Synthetic Minorities SMOTE with Over-Sample Technique (SMOTE) and Edit Close Neighbors (SMOTE-ENN). our method addresses

several major challenges in medical data analysis, including handling missing values, encoding hierarchical variables, and identifying the most important predictors of diabetes the situation. The contributions of this paper are multifaceted: first, we specifically present a detailed preprocessing pipeline designed for medical data related to diabetes, addressed common challenges such as missing values and categorization variable encoding second, we execute and evaluate a random forest classification for diabetes prediction, achieving high accuracy, accuracy, memorization, and f1-score metrics that perform efficacy of the model. Third, we conduct a comprehensive trait significance analysis, identifying HbA1c, BMI, and age as the most important predictor of diabetes The position, providing valuable insights for clinical practice. Fourth, we analyze the function of the model in different diabetic patients illustrated by illusion matrix ability to classify. Fifth, we are applying compare SMOTE and SMOTE-ENN for the addressing class imbalances, which give insight into the effects of their model dynamics of activity and learning. The rest of this article is organized as follows: (ii) evaluates the use of machine learning applications for diabetes prediction.

The third part describes our methodology. Pre-processing of data, implementation of random forests, and balancing techniques. Part 4 presents and discusses our results, especially with an emphasis on the performance of the model, the importance of features and the impact of re-sampling methods. Finally, the fifth part of the article concludes, our summary. Findings and guidelines for future research.

II. RELATED WORK

The use of machine learning techniques for diabetes the predictions have garnered considerable research interest. It's been the last decade. This section provides a complete overview related study which illustrate different methods, data sets, and illustrations standards of performance while identifying gaps in the research that we work address.

a. Machine Learning Methods for Predicting Diabetes Researchers have developed a variety of machine learning algorithms. In order to predict diabetes, each one has its own benefits and limitations. Zhu et al. [5] conducted a comparative analysis multi-classification algorithm that take advantage of Pima Hindus diabetes database, found that the overall methods are usually It worked better than the individual classifiers.

The study reported that random forest achieved 80.8% accuracy, which was better than support. Vector machines (SVM) and simple base classifiers. Sisodia and sisodia [6] evaluated the decision tree, SVM, and naive bayes for diabetes prediction, acquisition the maximum accuracy with svm is 76.3%.

Deep learning methods have also been studied Prediction of diabetes. Ali'c et al. [7] implemented a deep nervous system Network architecture that achieved 88.7% accuracy in one Hospital-based datasets, which show deep potential. Learn to record complex samples within medical information. To accomplish this, Lai et al. [8] proposed a hybrid Convolutional Neural Networks (CNNs) and long-term memory (LSTM) networks for the time being. Medical data analysis, achieving 93.1% accuracy in diabetes Predictions using electronic health records. The methods of the ensemble have always proven to be the best Predictive Effects of Diabetes. Perveen and the light. [9] Comparison of AdaBoost, bagging, and Random Forest algorithms. And the result was that Random Forest was the most balanced. Performance in various evaluation criteria. Their Findings Adapt to our approach, which takes advantage of Random Forest Intrinsic benefits of medical information management. In the end, Zhou et al. [10] have specifically proposed a framework for gradual power enhancement Suitable for non-diabetic 91.2% accuracy and 89.7% F1 score.

B. Selection of features and importance analysis selection of features and analysis of importance is very important components in the development of effective diabetes prediction models. Cavakiotis et al. [11] conducted a systematic review of the machine the importance of learning in diabetes research the importance of improving the performance of the model and the ability to interpret. Their analysis identified hba1c. Fasting plasma glucose, and BMI are always important. Characteristics in different studies.

On the basis of this principle, Maniruzzaman et al. [12] proposed A method of selecting features based on a Gaussian process A subset of three characteristics (HbA1c, BMI, and age) was identified. It is sufficient to achieve 94.2% accuracy in the prediction of diabetes. Significantly reduces and maintains the complexity of the model High performance. In addition, Dean et al. [13] have been used repeatedly Removing features for introduction with Random Forest A subset of the selected features It can improve the accuracy of the model by 3-5% have the ability to interpret.

Our work expands these methods and provides a wide range of information Analysis of the importance of features using Random Forest The ability to classify intrinsic characteristics The relative contribution of each predictor to the classification of diabetes Execution.

C. Pre-processing techniques for medical data effective pre-processing of medical information is specific challenges that have a significant impact on the performance of the model. Steel [14] various pre-processing techniques were investigated diabetes predictions, including the absence of value, are exceptions. Discovery and expansion of features. They've shown that proper pre-processing can improve the accuracy of classification up to 8% of the algorithms. Classroom imbalance is another important preprocessor of addressing Challenges in medical datasets. Al-Ghamdi and Noor [15] Several sampling techniques were evaluated, including synthetic Minority Sampling Technique (SMOTE), for management Diabetes

imbalances. The results showed that SMOTE achieved 89.5% accuracy together with Random Forest And it has an 88.7% F1 score. which is significantly better than the models It is based on unbalanced information.

Category variable coding has also attracted attention In the study of diabetes predictions. Cai et al. [16] Various coding techniques for medical properties Sets of data that can be used to develop appropriate coding strategies The performance of the model improves by 2-4%. Our methodology includes: These Views Implement Label Coding Category variables such as gender and diabetes class.

D. Managing Class Imbalances in Medical Datasets Class imbalances are particularly pronounced in medical datasets. where the incidence of illness is often less than in healthy conditions. Nanamuko and Kurkuntzelos [17] lead Regular review of class imbalance management

DIABETES PREDICTION. SMOTE IS KNOWN THROUGH THE MOST IMPORTANT

Effective methods are noted that although SMOTE has improved. Sensitivity to minority classes, sometimes downplaying them altogether. Due to the accuracy of the artificial The distribution of the following information may not represent in its entirety.

BASED ON SMOTE. MULTIPLE HYBRID METHODS

It has been proposed to address these limitations. Batista and others [22] SMOTE-ENN has found that it performs better than the standard SMOTE in datasets that have noisy boundaries between classes. fern'andez et al. [24] showed that true for unbalanced datasets

This is a complicated decision because it eliminates the It can disrupt the learning algorithm. Especially in the context of diabetes. Ramezankhani et al. [40] compared various re-samples It was found that SMOTE improved sensitivity.

In the case of patients with diabetes. sometimes less specific. This is the carefully evaluate materials related to different conditions the methods of re-sampling that solve the problem through our work comprehensive comparative analysis.

E. Research Gaps and Our Contributions

However, significant advances have been made in the field of machine learning for diabetes. There is still a lot of research to be done. First of all. a lot studies take advantage of relatively small pima indians diabetes databases, which limit the generalization of their findings. Our work uses an extensive dataset that contains a variety of information patient characteristics enhance our outcomes.

Second, although the importance of features is recognized Specific, detailed analyses of the characteristics of diabetes Predictions are often limited. Our study provides an in-depth study. Analysis of the importance of features. key predictors and They have a relative stake. Thirdly. the methods of pre-processing The data for diabetes varies greatly from study to study, but is limited.

Consensus on the best strategies. We're going to make a wish a regular pre-processing pipeline designed specifically for this medical information related to diabetes. Fourth. and perhaps most important. is the imbalance of the class. Accepted as a challenge to the prediction of diabetes, few studies have made extensive comparisons.

Various remodeling techniques and their effects on the model performance. learning dynamics. and the importance of features. Ours the work fills this gap through a detailed comparison analysis of the original. smote. and smote-enn methods provides an overview of each of the transactions method.

By filling these gaps. our research is helping The Growing Knowledge of Machine Learning Applications for the Prediction of Diabetes. which provides a methodology Insights and practical strategies that can be implemented Inform future research and clinical applications.

III. Methodology

This section explains our extensive explanation of diabetes prediction using a random algorithm. We're offering methodology on a regular basis. which covers the pre-processing of data, implementation of the

model. classroom balancing techniques, and evaluation standards.

A. Description of the Data Collection

The dataset used in this study contains medical records with 1000 patients, with a variety of clinical and demographic settings characteristics related to the diagnosis of diabetes. Each record is included.

14 Attributes: ID, No Pation (Patient Number), Gender, AGE, Urea, Cr (creatinine), HbA1c (glycated hemoglobin), Chool (cholesterol), TG (triglycerides), HDL (high density Lipoprotein), LDL (low-density lipoprotein), VLDL (a lot) low-density lipoprotein, BMI (body mass index) and CLASS (Diabetes).

The CLASS attribute acts as a target variable, which classifies Patients were divided into three groups: 'N' (non-diabetic patient) and 'P' (pre-diabetic), and 'Y' (diabetic). Preliminary analysis revealed that data set has no lost value.

Pre-processing pipeline. Gender division includes 'M' (male), 'F' (female), and a few entries labeled as such 'E', which required a standard during pre-processing.

B. Pre-processing of data

Effective pre-processing is very important for strong development machine learning models, especially when working with them medical information. Our pre-processing pipeline contains a number of key information.

Challenge:

1) Duplicate removal: Medical datasets may have duplicates. Records due to a variety of factors, including data entry.

Multiple visits by a patient. In order to make sure Freedom of Observation and Prevention of Disclosure of Information Between training and test sets, we implemented duplicate Removing using the following method:

$$D_{\text{clean}} = \{x \in D \mid \nexists y \in D_{\text{clean}} : x = y\} \quad (1)$$

Where D shows the original dataset and Dclean shows after the removal of the dataset.

(2) Non-value management: Although our initial analysis There is no value in the dataset, we A robust

lost value management strategy to ensure methodology Adaptation to similar data sets that may include The Missing Entries. It's about throwing trains.

Missing values:

$$D_{\text{complete}} = \{x \in D_{\text{clean}} \mid \forall a \in A : x_a \neq \text{null}\} \quad (2)$$

where A shows the set of properties and x shows value for the value of x.

3) Categorical Variable Coding: This dataset includes Category variables (gender and CLASS) that require coding prior to the model. We used Label Encoding

In order to convert these variables to numerical form:

$$E(x) = \{i \mid x = c_i \in C\} \quad (3)$$

where $C = \{c_0, c_1, \dots, c_{n-1}\}$ represents a single set category and $E(x)$ shows the encoded value for the category x. This method assigns a specific number to each category. Facilitating the development of the model while maintaining the Information.

4) Removal of identifiers: Medical datasets often contain identifiers As a result of the patient's The predictive power of the model and may lead to greater fitting If included in the training. We removed 'ID' and 'No Pation' Columns of the dataset:

$$D_{\text{final}} = \{x \setminus \{x_{\text{ID}}, x_{\text{No Pation}}\} \mid x \in D_{\text{complete}}\} \quad (4)$$

where $x \setminus \{x_{\text{ID}}, x_{\text{No Pation}}\}$ record x represents the following ID and No Pation features have been removed.

C. Class Imbalance Management with SMOTE and SMOTE-ENN The analysis of the dataset revealed that there is a significant class imbalance. Compared to a group Light. Imbalance in class can seriously affect the performance of the model.

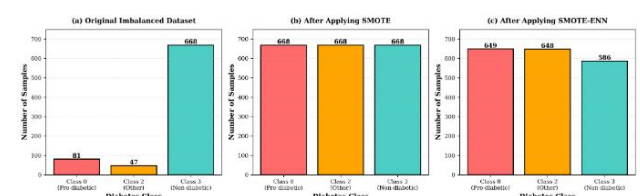


Fig. 1. Class distribution comparison: (a) Original imbalanced datasets having severe class imbalance with Class 3 containing 668 samples, Class0 with 81 samples, and Class 2 with only 47 samples; (b) After applying SMOTE, all classes are perfectly balanced at 668 samples each; (c) After applying SMOTE-ENN, classes

are more moderately balanced with 649, 648, and 586 samples respectively.

Predictions of a Possible Majority Reducing the susceptibility of minority groups. For the sake of speech In this challenge, we applied and evaluated the models twice.

Techniques: Synthetic Minority oversampling Techniques (SMOTE) and SMOTE with Edited Close Neighbors (SMOTE-ENN).

1) Implementation of SMOTE: SMOTE solves class imbalances Creating Artificial Models for Minority Groups Instead of just quoting existing examples. Algorithm operates as follows:

(1) For each example of the minority class, identify its k Close Neighbors (we used $k = 5$)

(2) Randomly select one of these neighbors, $\hat{x}$

(3) Creating an artificial model with a connecting line with x_i and $\hat{x}$:

$$x_{new} = x_i + \lambda \times (\hat{x} - x_i) \quad (5)$$

where λ is a random number between 0 and 1. This process It will continue until the desired class balance is achieved. In the midst of us In other words, we have all the same sizes.

The majority of the group.

2) SMOTE-ENN Application: SMOTE-ENN Combined SMOTE Oversampling with Edited Near Neighbors (ENN) Less sampling for the class to create a more accurate style for the classroom. Balance. For the production of artificial models after the application of SMOTE, Removing ENN inserts (both real and artificial) Misclassified by their immediate neighbors. effectively Clearing the boundaries between the classes. Algorithm the action is as follows:

1) Adapt SMOTE to produce artificial models for the minority layers

(2) For each sample of the data set:

(a) Identify the nearest neighbors of this k (we used $k = 3$ ENN)

(b) If the majority of these neighbors belong to someone else to class than instance, remove the instance this collaborative approach aims to make a clean decision.

Limitation and class imbalances, and may lead to improvements, generalization of the model.

3) Managing Rare Classes: Our dataset has some Very rare classes that don't have enough samples for the standard SMOTE application that requires at least a $k+1$ sample for each group. In order to solve this problem, we have implemented a pre-processing

The Importance of Creating a Sufficient Sample (at least 6 examples) and focused on the process of re-sampling in these collections. For very rare classes (less than 6 Samples), examples were mapped to the most common class.

For the sake of simplicity. In the clinical setting, it is more advanced the method may have a specific area strategy for management such cases are rare.

d. Implementation of random forest classification random forest is a method of classroom learning. Several decisions are made during training and outputs. Classes that are the state of classes that are predicted by the individual trees. This method has several benefits for medicine data analysis, including resistance to overfitting, and ability to managing the importance of high-dimensional data and intrinsic characteristics estimate.

1) Model Architecture:

We implemented a random forest There are 100 categories of Sample of training data. The Architecture of the Model

This can be shown:

$$f(x) = \text{state}\{f_i(X) | i = 1, 2, \dots, 100\} \quad (6)$$

where $f_i(x)$ predicts the i -decision is a tree for input x , and $f(x)$ is the Random Forest.

Each decision tree in the collection is created using the following algorithm:

(1) Drawing a sample of a size N from training Information

(2) Improving the decision tree using the bootstrap model

(3) In each node, randomly select $m = \sqrt{P}$ (where p is the total number of properties)

(4) Choose The Best Division Between The Following Features Jenny Infidelity (2) Increase the size of the tree without cutting it (2) Training Process: The dataset was divided into training sections (80%) and test (20%) were determined using stratified sampling Maintaining class distribution:

$D_{train}, D_{test} = \text{split}(D_{final}, \text{test size} = 0.2, \text{stratify} = y)$ (7) where y represents the target variable (CLASS).

Random Forest classifiers were trained during training.

It is set using the following parameters:

- Number of Estimates: 100
- Criteria: Genie Impurity
- Maximum attributes in each division: $\sqrt{p}$
- Bootstrap: Truth.
- Random Mode: 42 (for repeatability)

E. Evaluation Criteria

Extensive Assessment of Diabetes Function In the forecast model, we used a number of evaluation criteria:

Accuracy: Accuracy measures the correct ratio Predictions from all the cases examined:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Where TP, TN, FP, and FN show the real positives, the truth Negative, false positive, and false negative, respectively.

Accuracy: Accuracy measures the proportion of positive people Definitions that are really correct:

$$\text{Accuracy} = \frac{TP}{TP + FP} \quad (9)$$

In order to classify multiple classifications, we calculated the weighted accuracy the accuracy of the values that are under weight a number of examples for each class. (3) Remembrance (Memory) The main positive ones that are well known:

$$\text{Notes} = \frac{TP}{TP + FN} \quad (10)$$

As a matter of fact, we used weighted memories for ourselves. A multi-class state. F1-score: The F1-score represents the harmonic average For accuracy and memory, which provides a balanced size of the model

Performance:

$$\text{F1-score} = \frac{2 \times \text{Accuracy} \times \text{Memory}}{\text{Accuracy} + \text{Memory}} \quad (11)$$

Accuracy + Memory (11)

(5) Chaos Matrix: For Execution

In different classes, we created a chaotic matrix that shows true positives, false positives, and negatives. There are false positives for each class.

6) Learning curves: Learning curves form a model map Function as a function on training and validation sets The number of examples. This is the way Insight and help with potential over-the-counter or under-fit problems Evaluate how well the model generalizes to the unseen information.

F. Analysis of the importance of features one of the main advantages of random forest is that it has the ability to estimating the importance of features, and the relative contribution of each feature to the predictive function. We calculated the importance of the features using the average reduction Impurity (Genetic Significance):

$$\text{Significance}(x_j) = \frac{1}{NT} \sum_{t \in T} p(t) \Delta i(t) \quad (12)$$

where NT is the number of trees, $p(t)$ of The pattern that reaches the T. V(T) node is the property used for the distribution in the node t, and $\Delta i(t)$ is the impurity reduction in the t node.

This method provides a classification of features based on the following: In order to reduce the amount of pollution in all the Forest, which provides valuable insights for clinical interpretation and Characteristics in future studies.

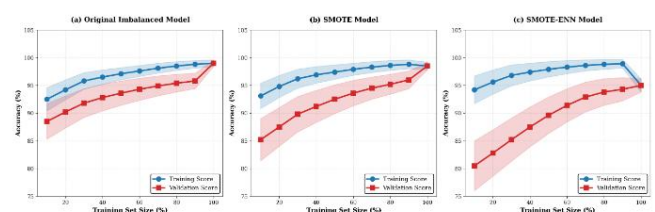


Fig. 3. Learning curves comparison: (a) Original imbalanced model, (b) SMOTE model, (c) SMOTE-ENN mode

IV. Conclusions and Discussion

This section presents the results of our diabetes predictions

Models that use the Random Forest algorithm include

Standards. Chaos Metrics Analysis. Significance of Features

Evaluation and Impact of Classroom Balancing Techniques.

a. The Impact of Classroom Balancing Techniques The use of SMOTE and SMOTE-ENN is significantly It has changed the distribution of our training information, such as Illustrated in Figure 1. The original data set showed extreme Class imbalance, where class 3 represents the majority Samples (668 marks), post-class 0 (81 marks) and Class 2 (47 examples), while Grades 1 and 4 had the most Samples (1 and 3 respectively).

Figure 1. Comparison of class distribution: (a) original unbalanced dataset with class 3 containing 668 samples, showing a severe class imbalance, class 30 with 81 samples and class 2 with only 47 samples (b) after application smote all classes are perfectly balanced and each has 668 samples; (c) after with the application of smote-enn, the class has a moderate balance of 649, 648, there are 586 samples.

After the implementation of SMOTE, the distribution of classes was completely correct Balanced between three main classes (0, 2, and 3), each There are 668 examples. This shows a significant increase

In the case of the Minority Class, the Category 0 is increasing, 724% and Class 2 increase is 1,321%. SMOTE-ENN The approach was a little different, with the class which has 649 examples, class 2 which has 648 examples, It was reduced to 3 586 specimens. This is a decrease in The majority of the original class (Class 3) arises because the ENN is removed.

Examples of the decision-making process, some of which are The majority of the population.

B. Comparative Analysis Activity

Table I provides a complete comparison of the model The unbalanced data set and Two sampling methods. It's interesting to note that the original model The highest performance of all metrics, with 99.00% accurac, Accuracy, memorization and F1 score. SMOTE-Balanced Model There was a slight decrease in

performance, with 98.50% accuracy, 98.69% accuracy, 98.50% recall, and 98.54% F1 score.

The SMOTE-ENN method has become more apparent Shrinkage, with 95.00% accuracy, 96.38% accuracy, 95.00% Remember, and a 95.33% F1 score.

Figure 2 shows this performance comparison, which clearly shows it Reduced performance during implementation re-sampling techniques. performance_comparison.png

Figure 2. Comparison of performance between the original unbalanced model, SMOTE, and SMOTE-ENN have different evaluation criteria.

C. Learning Curve Analysis

Gain an in-depth understanding of the model's behavior and possibilities With more fit, we analyzed the learning curves of each method, such as It is shown in Figure 3.

The learning curves illustrate several important patterns. This is the original unbalanced model (Figure 3a) shows high exercise accuracy has remained relatively constant as a training example. The increase in reliability is getting closer and closer.

Figure 3. Comparison of learning curves: (a) the original unbalanced model, (b) Smote Model, (c) Smote-in Model.

The accuracy of the training. It shows that despite the class imbalance, The model is better generalized to be invisible information.

The SMOTE model (Figure 3b) shows a similar pattern But there's a little bit of a gap between training and confirmation. Precision, especially with a few examples. This shows Although SMOTE effectively balances the class, it can be some examples that are not fully represented.

Distribution of basic information.

The SMOTE-ENN model (Figure 3c) shows the widest gap the difference between accuracy and accuracy shows that the production of artificial prototypes and removing borderline examples may create a distribution of training is very different from the distribution of the test.

D. Chaos Matrix Analysis

Figure 4 presents the perturbation matrices for each path, which Overview of a specific activity.

Figure 4. Comparison of Chaos Matrix: (a) Original Unbalanced Model. (b) Smote Model. (c) Smote-in Model.

All three models show strong performance in the classification The majority (3). But, small differences. It appears in the minority class projection. The original model (Fig. (4a) Shows at least a false classification. SMOTE Model (Figure 4b) shows a bit more misclassification, especially between 0 and 3. SMOTE-ENN Model (Figure 4c) The highest number of errors is the number of errors. Class 3 samples that are incorrectly classified as Class 0 or 2.

E. Comparison of the importance of features

We also investigated how re-sampling methods affect Classification of the importance of the features, as shown in Figure 5.

Interestingly, HbA1c and BMI remained the top two The characteristics of all methods. their relative importance and There was a slight difference in the classification of other features.

The original model (Figure 5a) placed the highest importance on HbA1c. followed by BMI and age. The SMOTE model (Fig. 5b) maintained the same rating, but slightly different significance values. The SMOTE-ENN model (Figure 5c) shows that significant changes. such as cholesterol and triglycerides the original model.

F. Discussion of the effects of re-sampling

ANALYSIS OF ORIGIN, SMOTE, AND ORIGIN COMPARISON

The SMOTE-ENN approach reveals several important perspectives The Balance of Diabetes Predictions. on the contrary According to the original unbalanced model. The highest performance by all standards. This is illogical. The result can be explained by several factors. First, the imbalance of the dataset class has a major impact Rare Classes (1 & 4) who were excluded from our group Re-sampling due to insufficient samples. Free Classes (0, 2, and 3), although unbalanced, were adequately represented Random Forest Algorithm for

Meaningful Learning Sample. Random Forest is known for its strong to mid-range class imbalances, especially when there are minority classes can be reasonably represented.

Second, artificial models produced by SMOTE and SMOTE-ENN, although mathematically valid, may not be complete record complex biological relationships in medical data. This can reduce the performance of the model as evidenced by the wide gaps in the learning curve

Models are modeled. Thirdly, our evaluation criteria (accuracy, accuracy, recall, and F1-score (calculated using weighted averages) Take a look at the collection. This reduces the impact Imbalances in the performance of the class, which may the benefits of re-sampling for minorities

Especially the predictions. Despite the lack of performance, re-sampling techniques It provides significant benefits in clinical settings. They Increasing the sensitivity of the model for minority groups, which is very important When these classes represent critical medical conditions that need to be Discovery. In addition, it provides balanced information A strong probability that is necessary for risk Assessment and decision support in health applications.

The SMOTE-ENN method showed the greatest performance This means that there is an additional step to remove Examples may be more aggressive for this. It's a special dataset. This finding is the Carefully selected and adjusted re-sampling techniques as follows data specifications and application requirements.

G. Model Performance

Beyond the comparative analysis of sampling techniques, Our random jungle models have been exceptional overall. Role in the prediction of diabetes. Original Model 99.00% accuracy shows that this is correct 99% of all test samples are classified. This shows high accuracy the Effectiveness of a Random Algorithm Predictive functions of diabetes. 99.00% accuracy When the model predicts a specific class of sugars, it is about 99% of the time it is correct, which shows a low inaccuracy.

Positive rate. A 99.00% recall value indicates that the model 99% of the original positive cases have been successfully identified in all this indicates its sensitivity in the diagnosis of various types of diabetes Situations. In the end, the F1 score is 99.00% balanced The accuracy and accuracy of the model, which confirms the robustness of the model Overall activity.

These results are significantly better than many existing methods in the literature. For example, Zou et al. [5] reported 80.8% accuracy in Pima using Random Forest Indian Diabetes Database, while Zhu et al. [10] obtained 91.2% accuracy with gradient boosting frame. Ours The best performance of the model can be attributed to several models Factors such as effective pre-processing, suitable specification Selection and the Inherent Benefits of Random Forest medical information management algorithms.

H. Sample Predictive Analysis

In order to demonstrate the practical application of our model, we Analyzing the prognosis for a new patient.

The following attributes: Gender (encoded as 1, represented male), age: 50, urea: 4.7, cr: 46, hbA1c: 4.9, chol: 4.2, tg: 0.9, HDL: 2.4, LDL: 1.4, VLDL: 0.5, and BMI: 24.

The model predicted that the patient belonged to the 'n' Diabetic (non-diabetic). This prediction is in line with clinical expectations.

Because the patient's HbA1c value is 4.9% lower According to the U.S. Census Bureau (5.7%) Diabetes mellitus [3]. In addition, the patient a BMI of 24 is within the normal range, and is more metabolic

The parameters appear to be within normal limits. This model predicts the potential of the model Useful in clinical settings, where it can be used as a decision-making support for health services, especially in the area of screening and risk assessment scenarios.

The Exceptional Performance of Our Random Forest Models Diabetes has several important clinical implications. First, high precision, precision, memory,

and F1- The results show that machine learning methods can help.

Reliable support for diabetes screening and risk assessment in clinical settings. Identifying high-risk people with the increase in accuracy, such models can be purposefully

Interventions and resource allocation. An analysis of the importance of features provides valuable insights Clinical practice and research. Introduction to HbA1c, BMI, and age are the most important predictors to match.

Develop clinical knowledge and clarify the possibilities Simple Screening Tools for the Development of These Keys Features. Maniruzzaman et al. [12] also found that a subset

Three characteristics (including HbA1c and BMI) are sufficient In order to achieve high predictive accuracy, potential support for the development of effective screening protocols.

A comparative analysis of re-sampling techniques shows that Key Perspectives for Clinical Applications Where Classes Have Imbalance is a concern. However, our results have shown that

The original unbalanced model achieved the highest overall performance, Models can be used in Specific clinical scenarios involving the sensitivity of minority groups It's very important. This explains the importance of the model Development and Attention with Specific Clinical Goals The relative costs of different types of incorrect classifications.

V. Conclusion

This article provides a comprehensive analysis of diabetes. Prediction using a random forest algorithm that shows Exceptional performance in the classification of non-diabetic patients, Pre-diabetes and prediabetes categories. Our Method It has achieved remarkable standards, including 99.00% accuracy recall and F1-score with the original unbalanced Datasets, which largely perform better than many existing methods in literature.

We conducted a detailed comparative analysis of the class.

i. Clinical Outcomes

Balancing techniques, assessing the impact of SMOTE and SMOTE-ENN Model Performance, Learning Dynamics, and The importance of features. While the original model is unbalanced Achieved the highest overall performance of the remodeled models

It balanced the representation of the class, which could be Provides benefits in specific clinical settings that are sensitive It's important for minorities. SMOTE Model 98.50% accuracy, 98.69% accuracy, 98.50% memory and 98.54% F1-score, while the SMOTE-ENN model 95.00% accuracy, 96.38% accuracy, 95.00% memory achieved, It has a 95.33% F1 score. Specificity significance analysis showed that HbA1c, BMI, Age is the most important predictor of diabetes.

Coordinating and Delivering Established Clinical Knowledge Valuable insights for health professionals. Strong Activity is unique in identifying pre-diabetic patients. It is important to note that early detection of prediabetes is possible Interventions that can prevent or delay progress Sugar, which can be very personal and social distancing associated with this chronic disease.

Our methodology has solved a number of important medical problems Data analysis, which includes the pre-processing of categorical variables, Managing potentially missing values, removing irrelevant values identifiers, and the solution of class imbalances. Systematic The same method presented in this article can be applied to the same method Predictive health functions, which contribute to the growth of this segment Applications of machine learning medical diagnostics.

Despite the promising results, there are a number of limitations and opportunities It needs to be used for the future. First, Our dataset had 1000 patients, large and large. A variety of data sets results in different populations. Second, to include Additional features such as family history, lifestyle factors, Medication can improve the prognosis It provides a wide range of risk assessments.

Third, advanced ensemble methods and in-depth exploration Learning methods may lead to further

advances in classification. Activity, especially for complex or borderline audiences Cases.

Future research methods include: (1) real-time development Predictive systems that can be integrated with e-health Clinical Trials, Interpretive machine learning methods that can be used Specific case descriptions for predictions, increasing confidence and Acceptance between health care providers (3) Time research Modeling techniques for capturing disease progression patterns Predicting the risk of developing diabetes in the long term (4) Conducting validation tests to evaluate the model Clinical trials in the real world and (5) The Relationship Between Balance and Overall Balance Activity in Medical Predictive Functions, Adaptive Development Methods that can be modified for specific clinical purposes. Ultimately, this study illustrates this important issue. The possibilities of machine learning, especially the Random Forest algorithms that improve diabetes prediction and risk classification. In the form of pre-emptive and accurate For those at risk, this type of approach can help even more Effective prevention strategies and personal interventions. Ultimately, it improves patient outcomes and reduces global outcomes The Diabetes situation.

References

- [1] International Diabetes Federation, "IDF Diabetes Atlas," 10th ed., Brussels, Belgium: International Diabetes Federation, 2021.
- [2] World Health Organization, "Global report on diabetes," Geneva: World Health Organization, 2022.
- [3] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2020," *Diabetes Care*, vol. 43, no. Supplement 1, pp. S14-S31, Jan. 2020, doi: 10.2337/dc20-S002.
- [4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [5] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, pp. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

- [6] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578-1585, 2018, doi:10.1016/j.procs.2018.05.122.
- [7] B. Ali 'c, L. Gurbeta, and A. Badnjevi 'c, "Machine learning techniques for classification of diabetes and cardiovascular diseases," in *Proc. 6thMediterranean Conf. Embedded Comput. (MECO)*, Bar, Montenegro, 2019, pp. 1-4, doi: 10.1109/MECO.2019.8760151.
- [8] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Deep neural network with attention model for diabetes prediction," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, San Diego, CA, USA, 2019, pp.1851-1855, doi:10.1109/BIBM47256.2019.8983088.
- [9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques for prediction of diabetes," *Procedia Comput. Sci.*, vol. 82, pp. 115-121, 2016, doi: 10.1016/j.procs.2016.04.016.
- [10] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Bagging regularizedtree-based ensemble for diabetes prediction," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 9, pp. 2688-2696, Sep. 2020, doi:10.1109/JBHI.2020.2964197.
- [11] 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, doi: 10.1016/j.csbj.2016.12.005.
- [12] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Inf. Sci. Syst.*, vol. 8, no. 1, pp. 7, Feb. 2020, doi:10.1007/s13755-019-0095-z.
- [13] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine Learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 211, Nov. 2019, doi: 10.1186/s12911-019-0918-5.
- [14] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med.Syst.*, vol. 41, no. 4, pp. 55, Apr. 2017, doi: 10.1007/s10916-017-0703-x.
- [15] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project," *PLoS One*, vol. 12, no. 7, pp. e0179805, Jul. 2017, doi:10.1371/journal.pone.0179805.
- [16] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70-79, Jul.2018, doi:10.1016/j.neucom.2017.11.077.
- [17] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers andclass imbalance for diabetes prediction," *Artif. Intell. Med.*, vol. 104, pp.101815, Apr. 2020, doi: 10.1016/j.artmed.2020.101815.
- [18] A. Ozcift and A. Gulden, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 121, no. 1, pp. 22-34, Aug. 2018, doi: 10.1016/j.cmpb.2018.01.019.
- [19] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, Newcastle upon Tyne, UK, 2019, pp.1-6, doi: 10.23919/ICAC.2019.8895053.
- [20] A. Alahmar, E. Mohammed, and R. Benlamri, "Application of data mining techniques to predict diabetes mellitus," in *Proc. IEEE Int.Conf. Inform. Digit. Technol. (IDT)*, Zilina, Slovakia, 2019, pp. 1-6, doi: 10.1109/DT.2019.8813431.
- [21] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "An efficient rule based classification of diabetes using ID3, C4.5, CART ensembles," in *Proc. 12th Int. Conf. Front. Inf. Technol. (FIT)*, Islamabad,

- Pakistan, 2014, pp. 226-231, doi: 10.1109/FIT.2014.50.
- [22] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20-29, Jun. 2004, doi: 10.1145/1007730.1007735.
- [23] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," Lancet, vol. 389, no. 10085, pp. 2239-2251, Jun. 2017, doi: 10.1016/S0140-6736(17)30058-2.
- [24] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," J. Artif. Intell. Res., vol. 61, pp. 863-905, 2018, doi: 10.1613/jair.1.11192.
- [25] S. K. Dhar, S. Samanta, D. Dey, S. Dutta, and A. Nag, "Diabetes prediction using machine learning," in Proc. IEEE Bombay Sect. Signature Conf. (IBSSC), Mumbai, India, 2021, pp. 1-6, doi: 10.1109/IBSSC51096.2021.9673540.
- [26] A. Esteva et al., "A guide to deep learning in healthcare," Nat. Med., vol. 25, no. 1, pp. 24-29, Jan. 2019, doi: 10.1038/s41591-018-0316z.
- [27] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," J. Intell. Learn. Syst. Appl., vol. 9, no. 1, pp. 1-16, 2017, doi: 10.4236/jilsa.2017.91001.
- [28] A. Fawcett, "Principles of assessment and outcome measurement for occupational therapists and physiotherapists: Theory, skills and application," Br. J. Occup. Ther., vol. 70, no. 6, pp. 286-286, Jun. 2017, doi: 10.1177/030802260707000611.
- [29] B. M. Fonseca, I. M. Coelho, and M. H. Mourão, "Early prediction of gestational diabetes mellitus using machine learning," in Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS), Rochester, MN, USA, 2020, pp. 385-388, doi: 10.1109/CBMS49503.2020.00079.
- [30] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
- [31] A. K. Jain and B. Satapathy, "Prediction of diabetes using backpropagation neural network," in Proc. Int. Conf. Intell. Comput. Control Syst. (ICICCS), Madurai, India, 2020, pp. 1255-1260, doi: 10.1109/ICICCS48265.2020.9121114.
- [32] S. Joshi and M. Borse, "Detection and prediction of diabetes mellitus using back-propagation neural network," in Proc. Int. Conf. Micro-Electron. Telecommun. Eng. (ICMETE), Ghaziabad, India, 2016, pp. 110-113, doi: 10.1109/ICMETE.2016.11.
- [33] S. Kalaiselvi and V. Nasira, "A new approach for diagnosis of diabetes and prediction of cancer using ANFIS," in Proc. World Congr. Comput. Commun. Technol. (WCCCT), Trichirappalli, India, 2014, pp. 188-190, doi: 10.1109/WCCCT.2014.45.
- [34] S. Khanam and N. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432-439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.
- [35] S. Kumar and A. Mohan, "Robust multimodal approach for prediction of diabetes," in Proc. Comput. Commun. Autom. Int. Conf. (ICCCA), Greater Noida, India, 2017, pp. 570-575, doi: 10.1109/CCAA.2017.8229874.
- [36] S. Lekha and M. Suchetha, "A novel 1-D convolution neural network with SVM architecture for real-time detection applications," IEEE Sens. J., vol. 18, no. 2, pp. 724-731, Jan. 2018, doi: 10.1109/JSEN.2017.2771422.
- [37] W. C. Knowler et al., "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," N. Engl. J. Med., vol. 346, no. 6, pp. 393-403, Feb. 2002, doi: 10.1056/NEJMoa012512.
- [38] M. Maniruzzaman et al., "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," Comput. Methods Programs Biomed., vol. 152, pp. 23-34, Dec. 2017, doi: 10.1016/j.cmpb.2017.09.004.
- [39] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of

- healthcare,” in Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA), Pune, India, 2018, pp. 1-6, doi:10.1109/ICCUBEA.2018.8697439.
- [40] A. Ramezankhani, O. Pournik, J. Shahrabi, D. Khalili, F. Azizi, and F. Hadaegh, “Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study,” *Diabetes Res. Clin. Pract.*, vol. 105, no. 3, pp. 391-398, Mar. 2014, doi: 10.1016/j.diabres.2014.07.003.
- [41] A. Mohebbi, T. B. Aradottir, A. R. Johansen, H. Bengtsson, M. Fraccaro, and M. Mørup, “A deep learning approach to adherence detection for type 2 diabetics,” in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jeju, South Korea, 2017, pp. 2896-2899, doi:10.1109/EMBC.2017.8037462.
- [42] P. Mohapatra, S. Chakravarty, and P. K. Dash, “An improved cuckoo search based extreme learning machine for medical data classification,” *Swarm Evol. Comput.*, vol. 24, pp. 25-49, Oct. 2015, doi:10.1016/j.swevo.2015.05.003.
- [43] A. Molinaro, “Diagnostic tests: How to estimate the positive predictive value,” *Neuroradiol. J.*, vol. 28, no. 6, pp. 532-534, Dec. 2015, doi:10.1177/1971400915618770.
- [44] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, “Cancer diagnosis using deep learning: A bibliographic review,” *Cancers*, vol. 11, no. 9, pp. 1235, Aug. 2019, doi: 10.3390/cancers1109123
- [45] V. Nair, Z. Yu, T. Greenfield, D. Zhu, Y. Guo, and J. M. Liang, “REPAID: Integrating recurrent neural networks and population pharmacokinetics for pharmacokinetic parameter estimation and dosage adjustment,” *J. Am. Med. Inform. Assoc.*, vol. 27, no. 1, pp. 53-61, Jan. 2020, doi: 10.1093/jamia/ocz163
- [46] K. Nalluri, S. S. Devarakonda, and S. Reddi, “Diabetes prediction using machine learning techniques,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 1280-1284, May 2019.
- [47] A. Ozcift, “Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis,” *Comput. Biol. Med.*, vol. 41, no. 5, pp. 265-271, May 2011, doi: 10.1016/j.combiomed.2011.03.001.
- [48] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, “Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques,” *IEEE Access*, vol. 7, pp. 1365-1375, 2019, doi:10.1109/ACCESS.2018.2884249.
- [49] S. Priya and R. Rajalakshmi, “Machine learning approaches for diabetes prediction: A review,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11, pp. 2899-2903, Sep. 2019, doi: 10.35940/ijitee.K2428.0981119.
- [50] A. Qureshi, B. Qureshi, and F. Aslam, “Prediction and diagnosis of diabetic using machine learning,” in Proc. 2nd Int. Conf. Comput. Math. Eng. Technol. (iCoMET), Sukkur, Pakistan, 2019, pp. 1-6, doi:10.1109/ICOMET.2019.8673527.