Comparative Analysis of Resampling Techniques and Machine Learning Classifiers in Multiclass Classification of Diabetes Mellitus

Afshan Hashmi
Department of Computer Science & Engineering
Jamia Hamdard
New Delhi, India
afshanhashmi786@gmail.com

Md Tabrez Nafis
Department of Computer Science & Engineering
Jamia Hamdard
New Delhi, India
tabrez.nafis@gmail.com

Sameena Naaz
Department of Computer Science & Engineering
Jamia Hamdard
New Delhi, India
snaaz@jamiahamdard.ac.in

Imran Hussain
Department of Computer Science & Engineering
Jamia Hamdard
New Delhi, India
hussain.imran@gmail.com

2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-0085-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSAS57918.2023.10331822

***Abstract*— This research study explores the effects of various resampling techniques with different machine learning classifiers on the accuracy of multi-class classification of Diabetes using an imbalanced dataset. The diabetes dataset of Mendeley is a multi-class dataset with information about patients with no diabetes, pre-diabetes, and diabetes. The dataset is imbalanced, where the majority class is diabetic. This study is a comparative analysis of various oversampling techniques, undersampling techniques, and hybrid techniques with different machine learning algorithms to accurately classify the person as diabetic, pre-diabetic, or non-diabetic. Eight machine-learning algorithms and ten resampling techniques were applied to the dataset to classify the patient accurately. The result indicates that the combination of XGBoost with K mean smote and smote N attains the highest accuracy of 99.2%. It also suggests that oversampling techniques perform better than undersampling techniques and hybrid techniques.**

Keywords*— *Multiclass Classification, Resampling, Smote, Diabetes Mellitus, Imbalanced Data.

I. INTRODUCTION

A significant number of the global population has been impacted by the chronic condition known as diabetes mellitus. The most noticeable indication of this condition is an increase in blood glucose levels. Diabetes that is not properly managed can result in major health problems such as diabetic retinopathy, heart disease, renal failure, and stroke [1]. Pre-diabetes was first considered in 1997 as an intermediate diagnosis by the expert committee on Diagnosis and classification of Diabetes Mellitus because it indicates the high probability of the development of diabetes in the future [5]. It has been reported that 5–10% of patients with untreated prediabetes go on to develop diabetes [6, 7].

Sometimes unawareness of diabetes-related information and symptom can turn the prediabetic person into a diabetic. So, early detection of prediabetes is very crucial and this study includes the prediabetic class also. Scientists contend that both inherited and environmental factors affect the development of diabetes. Reduced disease-related consequences and risk factors can be achieved with early detection and treatment using medical data [2]. Medical data mostly comprises unbalanced datasets. Data that are out of balance tend to have an uneven distribution of samples among the different classifications. Medical data is uneven, which makes it challenging to extract different resources. Even though machine learning has come a long way, designing effective algorithms that rely on uneven data remains a challenging task [3]. Machine learning has been proven a significant tool for the early prediction and diagnosis of diabetes but the imbalance in the dataset can lead to biased accuracy. A machine learning algorithm might give a high false negative rate when an imbalanced dataset is used. The imbalanced dataset is those whose majority class has a much higher number of samples than the minority class. Various resampling techniques for instance undersampling, oversampling or hybrid techniques can be used to address the problem of imbalance in the dataset. Undersampling is used to decrease the majority class samples and oversampling is used to increase the minority class samples. The number of samples in the majority class is decreased by undersampling, whereas the number of samples in the minority class I is increased by oversampling.

Increased complexity, extended execution time, and overfitting might occur due to oversampling on the other hand undersampling create a risk of losing important information in the process of reducing the number of the majority class [4,5]. In this study, the majority class is the people with no diabetes while the other two classes, diabetic and prediabetic are the minority class. So to balance the datasets ten resampling techniques have been used namely, smote, SVM-SMOTE, borderline smote, k means smote, smote N, smote ENN, smote Tomek, ADASYN, random oversampling, and random undersampling. Eight machine learning classifiers namely, linear discriminant analysis, XG boost, random forest, logistic regression, Gaussian Naive Bayes, decision tree, and support vector classifier were applied to Mendeley's diabetes dataset for the multi-class classification.

A. Purpose

The purpose of this research is to investigate the effects of various resampling techniques combined with different machine learning classifiers on the accuracy of multi-class classification of diabetes using an imbalanced dataset. To overcome class imbalance, the study compared the performance of several resampling techniques, including oversampling, undersampling, and hybrid approaches. The research also attempted to assess how well different machine learning algorithms performed in correctly categorizing people as diabetic, pre-diabetic, or non-diabetic. The ultimate objective was to compare the effects of various resampling techniques on balanced and imbalanced datasets, for enhancing diabetes categorization accuracy, which can have major effects on early detection, individualized care, and enhanced patient care. This research paper is structured into 6 parts, after this section, Section 2 consists of the literature survey. Section 3 is the methodology that provides a thorough explanation of the experimental approach. Section 4 presents a detailed analysis of the result and major research findings, Section 5 is Discussion, which, consist of research implication and practical implication, and the conclusion and future work are presented in Section 6.

II. Related Work

Early identification or detection is the most crucial factor for preventing or managing Diabetes. It can be diagnosed and prevented with the use of an intelligent system based on illness symptoms and lab testing A lot of research has been carried out for this using machine learning techniques and a lot of researchers has achieved decent result in terms of accuracy but a high level of accuracy might be of no use if the dataset is unbalanced so various studies had tried to balance the data using data augmentation or resampling techniques to get an unbiased maximum accuracy in detecting predicting diabetes. Some of the related works are as follows:
Research done by [8] proposes the use of both oversampling and undersampling techniques to improve the performance of six machine learning algorithms. The author used the Pima India dataset and evaluated the model on the accuracy, precision, recall, and F1 measure for the performance of binary classification.

In [9] SmoteTomek was used for balancing the Pima India Dataset and six machine learning algorithms including Artificial Neural Network were applied Shapely Additive Explanation for explaining was used for each model.

Techniques of Smote and Tomek links were used in [2] and the result of the decision tree, Random Forest, and Support vector machine were compared both with and without balancing and the best accuracy obtained was 92.2%.

The authors did a comparative study between machine learning algorithms and oversampling techniques and suggested that Ada boost algorithm and SVM Smote have better performance[10]. In [11] researchers found that Deep Convolutional Neural Network along with SMOTE provided better results using a dataset of Pima India.

Artificial intelligence approaches have recently been used by researchers [12]to predict diabetes. A novel SMOTE-based deep LSTM system was subsequently created to identify diabetes early. This approach manages the diabetes dataset's class imbalance, and its prediction accuracy is evaluated. In this research, a SMOTE-based deep LSTM strategy for diabetes prediction is proposed along with analyses of CNN, CNN-LSTM, ConvLSTM, and deep 1D-convolutional neural network (DCNN) methodologies. When the accuracy of the suggested model was evaluated using the diabetes dataset, the proposed technique had the greatest prediction accuracy, 99.64%. Authors in [13] tried to improve the performance of Random Forest by balancing the dataset obtained from Kaggle with Smote TomekLink and got an accuracy of 86%. With the use of several strategies, including the weighted class approach, SMOTE (and its variations), and a straightforward artificial neural network model as the classifier, the authors tried to address the issue of class imbalance [14]. The authors of [15] use SMOTE and ROSE methods to balance the PIMA dataset and concluded that the random forest achieves the highest F Score of 0.83. A Multivariate Imputation Chain Equation was used for data pre-processing. The class imbalance issue has been handled in [16] using the SMOTE and ADASYN techniques. To find out which algorithm gives the best results, the authors explored machine learning classification methods, such as decision trees, SVM, Random Forest, Logistic Regression, KNN, and different ensemble approaches. The suggested system produced the best result in the XGBoost classifier with the ADASYN technique with 81% accuracy after training on and evaluating all of the classification models.

A. Originality

The following gaps were observed in the related research done so far:

- Most of the work has been done on the binary classification of diabetes datasets and
- SMOTE has been the most common technique used for class imbalance.
- Most of the research done in this area has used only oversampling techniques
- The maximum number of resampling techniques used in the research so far is six and all of them were oversampling techniques

This study is novel in the following ways:

- It has used ten resampling techniques on eight machine-learning classifiers
- It has dealt with the multi-class classification problem.
- It has not only used oversampling techniques but also Undersampling and hybrid techniques.

III. Methodology

A. Dataset

The dataset used in this research is the Mendeley Diabetes dataset which is a multi-class classification dataset that contains information about patients with diabetes, a patient without diabetes, and a patient with prediabetes. This dataset contains 1000 rows and 14 columns Each of the 14 features of this dataset has been described in **Table 1.**

TABLE I. Description of the dataset

Feature Name	**Description**
ID	Unique identifier for each record
No_Pation	Patient identification number
Gender	Male or female
Age	Patient age in years
Urea	Urea level in the blood of the patient
Cr	Creatinine ratio of the person
HbA1c	Level of glycated hemoglobin over the past 2-3 months
Chol	Cholesterol
TG	Triglycerides
HDL	High-Density Lipoprotein(good cholesterol)
LDL	Low-Density Lipoprotein(bad cholestrol)
VLDL	Very Low-Density Lipoprotein(carries triglycerides)
BMI	Body Mass Index is the body fat based on a person's weight and height.
Class	Diabetic, Prediabetic or non diabetic

where 844 samples belong to the "diabetic" class, 103 samples belong to the "non-diabetic" class, and only 53 samples belong to the "prediabetic" class As depicted in Fig. 1. It is clear that the classes in the dataset are imbalanced, with the majority class "diabetic" having a much higher number of samples and "prediabetic class having the least number of samples.

B. *Data Preprocessing*

The required pre-processing of the dataset was done as follows: This dataset has no missing values. The outliers were checked with the boxplot and it shows that the diabetic class has the most outliers but with the domain knowledge it was found that the values that tend to look outliers are possible values in the case of diabetes so it's not considered outliers in this study. as shown in Fig.2.

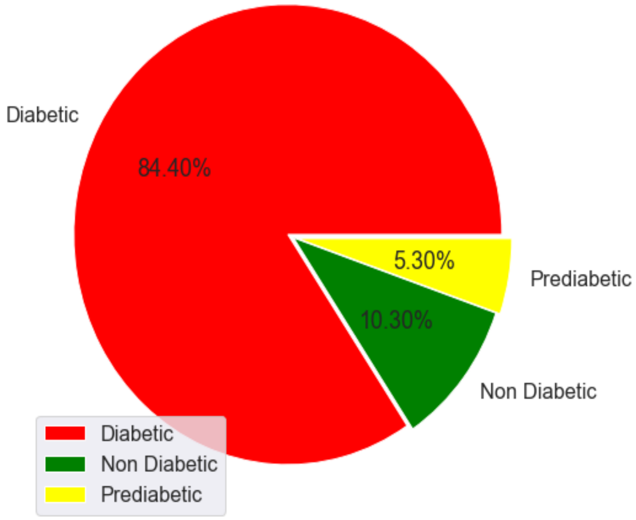

Fig. 1. Class imbalance in the dataset

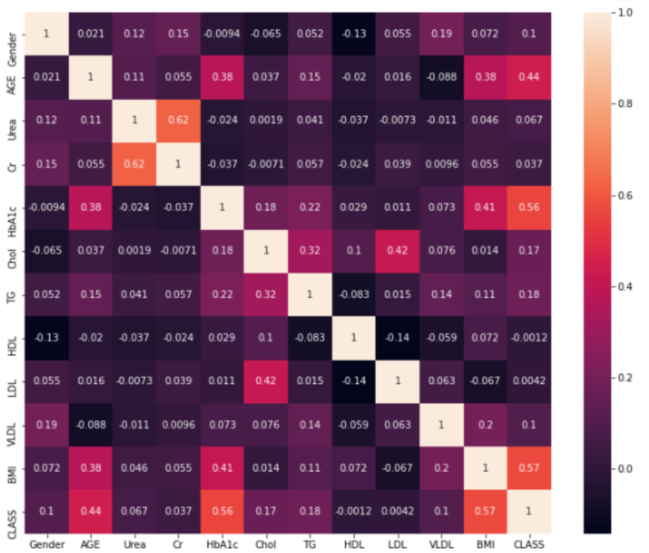

Fig. 2 Correlation Heatmap

Next, the correlation is checked if there is any in the dataset and it was found that Urea and creatinine have the highest correlation and BMI followed by HbA1c has the second and third highest correlation with the class. But still, the correlation is less than the threshold value so it hasn't been dropped them. However, ID and No. Of Pation cannot add any information regarding the Class of the patient so it has been dropped so now the dataset is left with 11 features and 1 label i.e. Class. Where Class 0 represents as Non-Diabetic, 1 is prediabetic and 2 is Diabetic. The dataset contains features that have different scales of value to normalize the data so that all the values scaled between 1 to 0 min-max scaler method has been used as defined by (1) below.

$$x = \frac{value - min}{max - min} \quad (1)$$

Where,
x is the scaled value
$value$ is the original feature value
min denotes the minimum feature value
max denotes the maximum feature value

C. *Data Resampling*

As mentioned before the dataset is imbalanced so to balance the dataset resampling techniques have been used that involve Oversampling of the minority classes, Undersampling of the majority classes, and hybrid techniques. Fig.3. depicts the no. of samples in each of the three classes before applying any resampling techniques. Figure 4 illustrates how undersampling effectively reduces the sample count within the majority classes, aligning it with the number of samples in the minority classes. Likewise, in Fig. 5, the effect of oversampling techniques is showcased, where the quantity of samples in the minority classes is elevated to match that of

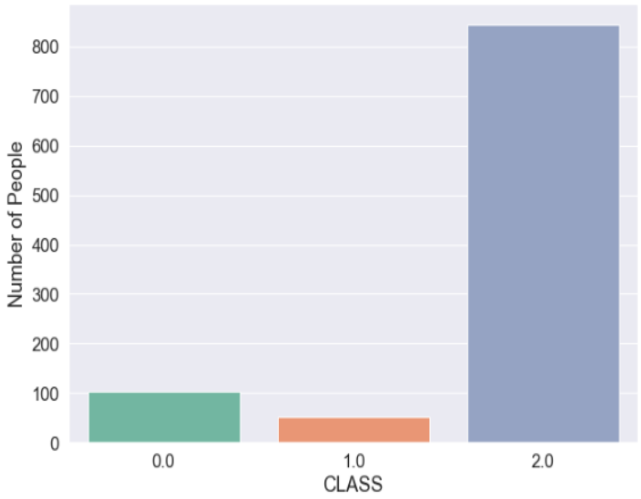

Fig. 1. No. of samples in each class without resampling

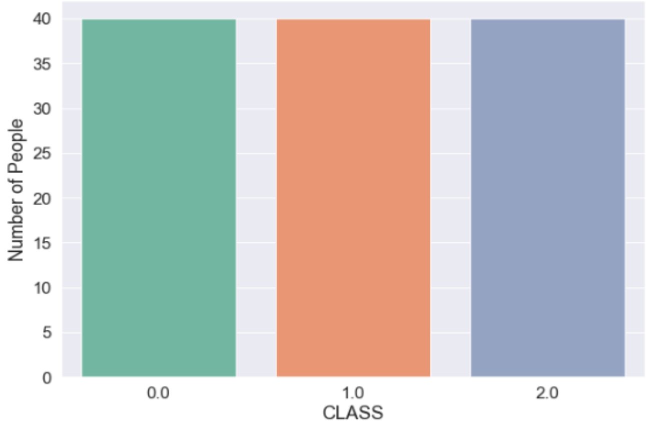

Fig. 2 No. of samples in each class after undersampling

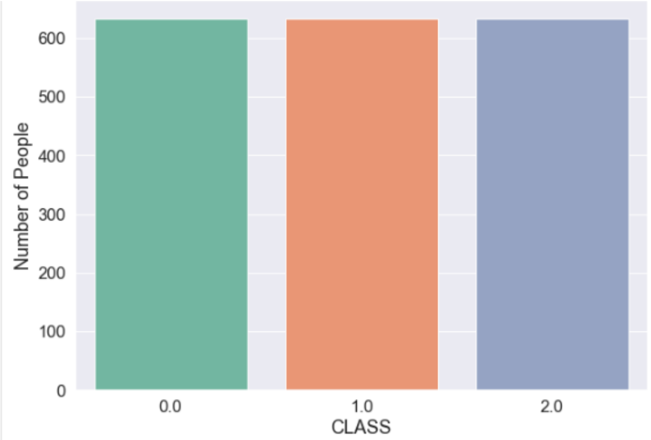

Fig. 3. No. of samples in each class after oversampling

the majority class.

The ten resampling techniques used in this study are as follows:

1) SMOTE means Synthetic Minority Oversampling techniques. It artificially creates some instances in the minority class to balance the class distribution in the dataset. The instance it creates is different from the other instance in that class. Thus obtaining generalization for the dataset.

2) ADASYN stands for Adaptive Synthetic Sampling it is also an oversampling technique that produces synthetic samples for the minority class in a way that prioritizes the more challenging minority samples to learn. It calculates the minority sample's density distribution and creates new samples in low-density areas. This ensures that ADASYN produces more synthetic samples in regions with the greatest class imbalance, enhancing the minority class's ability to be classified accurately.

3) Random oversampling as the name suggests randomly duplicates the instances of the minority class to increase the number of instances so that class distribution is balanced.

4) SVM SMOTE It is a combination of two methods SMOTE and SVM in which first smote is utilized to increase the data points and then SVM is used to categorize the data thus by adding both techniques together it can significantly improve the performance of a classification task.

5) Borderline SMOTE It focuses on borderline instances, which are very close to the majority class and are often misclassified by the classifiers. SMOTE is applied only to those borderline examples, generating synthetic examples only in the areas of the feature space where the class distributions overlap. This way, the synthetic examples generated are more likely to represent the true distribution of the minority class, and the learning algorithm can better generalize to new data.

6) K-Means SMOTE It is a data augmentation technique that combines the K-Means clustering algorithm with SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic data points for balancing an imbalanced dataset.

In this technique, the K-Means algorithm is used to cluster the minority class data points into K clusters. Then, the SMOTE algorithm is applied to these clusters to generate synthetic data points. The feature values of the minority class data points are combined with the feature values of their closest neighbors within the same cluster to create the new synthetic data points.

7) SMOTE N It is an effective technique for oversampling datasets with nominal features that are unbalanced. The first step of the SMOTE N algorithm is to identify each minority class sample's k nearest neighbors. Then, by interpolating between the k-nearest neighbors, it creates synthetic samples. To interpolate, a value between the two closest neighbors is chosen at random, and a new sample is then created using that value without disturbing the distribution of data.

8) SMOTE ENN It is a hybrid method for sampling skewed datasets. It combines the benefits of the data cleaning method Edited Nearest Neighbors (ENN) and the data augmentation method SMOTE. However, smote can add noise which affects the proper classification. By initially creating synthetic samples from the minority class using SMOTE, SMOTE ENN combines the benefits of SMOTE and ENN. The dataset is then cleaned up by using ENN to eliminate noisy samples. As a result, the dataset has become more balanced, and machine learning algorithms have become more accurate.

9) SMOTE Tomek means Synthetic Minority Over-sampling Technique Tomek Links It is a hybrid approach of data cleaning and data augmentation where smote is used for augmenting the data but while doing that it adds some noise as well. So, the Tomek link is used to get rid of that noise. It removes samples that can be quickly assigned to the majority class. Tomek Links are samples that are commonly found along the decision line between the majority and minority classes. By eliminating samples that are likely to be incorrectly identified, deleting Tomek Links can assist to increase the accuracy of machine learning systems.

10) Random undersampling It is a simple and effective way to balance imbalanced datasets. However, random undersampling can introduce bias into the dataset if the samples are not deleted randomly. It downsizes the sample of the majority class without adding noise to it.

After using the oversampling techniques all three classes got an equal percentage of distribution in the dataset as shown in Fig. 6.

D. Experimental setup

Jupyter Notebook was employed for the implementation using Python programming language. basic machine learning best practices is used to ensure the dataset's generalization and the model's ability to generate accurate predictions on additional, unknown data. The steps taken were as follows:

Data Splitting: After preprocessing and resampling the data, the data was split using a train-test split where 75% of the data was used as training data and 25% of the data was used for testing. the test set was kept separate and only utilized for assessment.

Model Training: the model is trainingon the training set using eight machine learning methods discussed later in this section. The model learned to detect patterns, characteristics, and correlations in the data during this training phase.

Model Evaluation: Following training, the model's performance Is assessed.

The choice of machine learning algorithms was based on several considerations, including the nature of the problem, the characteristics of the dataset, and the research objectives. The goal of this research is to perform a comprehensive comparison analysis of different resampling techniques to address class imbalance and effectively classify diabetes cases. To achieve this, a diverse set of widely-used machine learning algorithms is aelected , each known for its unique strengths and suitability for different types of data.

Eight selective machine-learning algorithms were used for modeling

1) Linear Discriminant Analysis (LDA): LDA is a statistical method that analyzes the data to find patterns and differences between classes. It seeks a linear combination of attributes that best differentiates various classes

2) XG Boost: XG Boost is an abbreviation for Extreme Gradient Boosting. It is a machine learning algorithm that uses an ensemble of weak prediction models (decision trees) to make accurate predictions. It is famous for being effective and having the capacity to manage complex data patterns.

3) Logistic Regression: is a classification method that forecasts the likelihood of a particular result. *It works by modeling the relationship between input features and the probability of a certain class using a logistic function.*

4) Random Forest: An ensemble learning technique called Random Forest uses several decision trees to generate predictions. *The final forecast is created by averaging all of the trees' predictions, each of which has been trained on a different subset of the data and attributes.*

5) Decision Tree: A Decision Tree is a structure resembling a flow chart in which each internal node corresponds to a feature or characteristic, each branch corresponds to a decision rule, and each leaf node corresponds to the conclusion or class label. It is flexible enough to be used both in classification and regression tasks.

6) Gaussian Naive Bayes: Gaussian Naive Bayes is a basic but powerful classification technique based on Bayes' theory. It assumes that the features are independent and follow a Gaussian distribution. This approach calculates the likelihood that a given data point pertains to a specific class, ultimately selecting the class associated with the highest computed probability.

7) K Nearest Neighbor (KNN): K Nearest Neighbor is a classification technique that allocates a single data point to a class based on the classes of its feature space nearest neighbors. The number of neighbors considered is determined by the value of K

8) Support Vector Machine: SVM is a very effective machine algorithm that can be used for regression as well as classification problems. It discovers a hyperplane in the feature space that separates various classes as much as possible. It can handle complicated patterns and works well in three-dimensional areas.

9) All eight machine learning algorithms were used for the classification of diabetic, pre-diabetic, and non-diabetic classes. Since Ten resampling techniques have been so it is tested on each resampled data with all the classifiers to

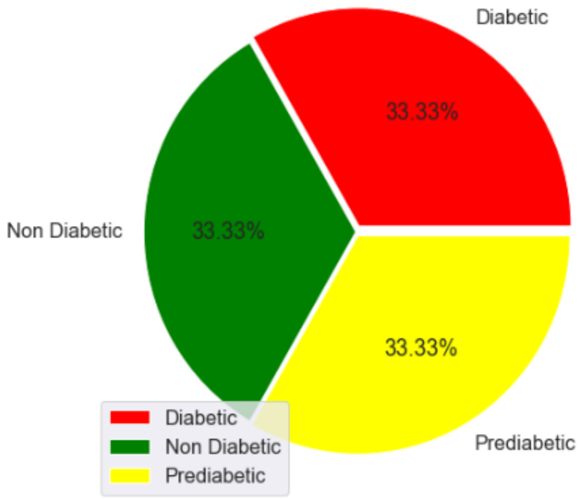

Fig. 4. Equal Percentage of classes after resampling

evaluate their performance. In this study accuracy(2) has been considered as the evaluation metric because the dataset has been balanced before modeling and in that case accuracy can be an effective performance metric. The For each algorithm, the classification error(3) is obtained by subtracting the accuracy from 1 This is because classification error is essentially the proportion of misclassified instances, and it's complementary to accuracy.

$$Accuracy = \frac{Number\ of\ correctly\ classified\ instance}{Total\ number\ of\ instance} \quad (2)$$

$$Classification\ error = 1 - Accuracy \quad (3)$$

IV. Result

The performance of various resampling techniques combined with different machine learning classifiers was evaluated using multiple evaluation metrics. Table 2 presents the classification accuracy achieved by each resampling technique and classifier combination. Among the resampling techniques, XGB (Extreme Gradient Boosting) consistently yielded the highest accuracy across all classifiers, with values ranging from 0.960 to 0.992. LDA (Linear Discriminant Analysis) on the other hand demonstrated the lowest performance, achieving accuracies between 0.876 and 0.904. Decision Tree obtained accuracy ranging from 0.988 to 0.964. Random Forest (RF) and Logistic Regression (LR) attained accuracy values ranging from 0.936 to 0.980 and 0.904 to 0.936, respectively.

Regarding the resampling techniques, Smote and Adasyn exhibited similar performance across most classifiers, with accuracies ranging from 0.888 to 0.900 and 0.896 to 0.920, respectively. Random oversampling and SVM Smote achieved accuracies ranging from 0.888 to 0.920 and 0.924 to 0.992, respectively. Borderline Smote and Kmeans Smote demonstrated comparable accuracies, ranging from 0.892 to 0.904 and 0.936 to 0.992, respectively.

SmoteN, Smote ENN, and Smote Tomek yielded accuracies ranging from 0.876 to 0.920, 0.912 to 0.960, and 0.836 to 0.956, respectively. Random undersampling showed the lowest accuracy values, ranging from 0.796 to 0.964. The Classification error of the used algorithms are as follows: LDA's error is 0.087, XGB's is 0.012, RF's is 0.024, LR's is 0.063, KNN's is 0.056, GNB's is also 0.056, SVC's is 0.063, and DT is 0.012.

It is important to note that apart from accuracy, future research should consider additional metrics, for evaluation such as precision, recall, and F1-score, to gain a comprehensive understanding of the performance of these resampling techniques in the context of the specific dataset and classifiers employed.

A. Major research findings

1. Oversampling was proved much better than Undersampling with all the ML algorithms as shown in Fig. 7.

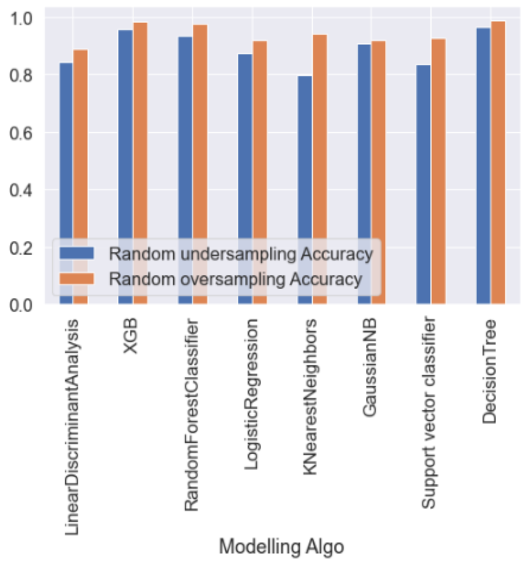

Fig. 7 Accuracy: Oversampling vs Undersampling

2. The accuracy was higher without using any resampling technique because use the dataset is highly imbalanced and a classifier can achieve high accuracy by simply predicting the majority class most of the time.
3. XGBoost (XGB), which regularly outperformed other machine learning algorithms by obtaining high accuracy across a range of resampling strategies, showed up as the best-performing one. It had one of the greatest accuracies for each resampling technique, demonstrating its potency in correctly identifying diabetes.
4. Random Forest (RF) and Logistic Regression (LR) also showed competitive performance, obtaining high accuracy levels consistently across various resampling strategies.
5. The performance of Linear Discriminant Analysis (LDA) was noticeably worse than that of other classifiers, indicating the need for new methods for the categorization of diabetes into many classes.
6. SMOTE, Adasyn, and random oversampling performed consistently better than random undersampling, proving the value of creating synthetic samples to correct the class imbalance.
7. The study addressed the necessity of using the right combination of resampling techniques and machine learning algorithms to achieve accurate diabetes classification since performance differed depending on the technique-algorithm pairing.
8. The findings suggested that using XGBoost, Random Forest, and Logistic Regression, together with appropriate resampling strategies, might improve diagnosis outcomes in real-world diabetes screening circumstances

The overall performance of the classifiers and the resampling technique is depicted in Fig. 8.

V. Discussion

A. *Research Implications*

The findings of this study have several implications for the field of multi-class classification of imbalanced datasets, specifically in the context of diabetes classification. The following implications can be drawn from the research:

1) Algorithm Performance:

- XGBoost (XGB) emerged as the top-performing machine learning algorithm, exhibiting the highest accuracy in classifying individuals as diabetic, pre-diabetic, or non-diabetic. This highlights the effectiveness of XGBoost in handling imbalanced datasets and its potential for accurate classification in the domain of diabetes diagnosis. For multiclass classification problems XGB by default uses the one vs rest strategy as this is the simple, computationally inexpensive, and preferred strategy for a moderate-size dataset.
- Linear discriminant analysis (LDA) was identified as the weakest performer among the algorithms tested. This suggests that LDA may not be well-suited for this particular classification task and alternative algorithms should be explored
- Execution time of LDA is 0.008 seconds, XGB is 0.302 sec, RF is 0.234 sec, LR is 0.031 sec, KNN is 0.00 sec, GNB is 0.00 sec, SVC is 0.015, DT is 0.013 sec. Based on the execution time and accuracy obtained by the algorithms DT appears to be particularly suitable because of its low execution time of 0.013 seconds and high accuracy of 98.8%

2) Resampling Techniques:

- When compared to undersampling techniques and hybrid approaches, oversampling strategies such as SMOTE (Synthetic Minority Over-sampling Technique), SVM Smote, K-means Smote, and Borderline Smote performed better. This suggests that for unbalanced datasets, producing synthetic samples to balance the class distribution is more helpful in boosting classification accuracy.
- Hybrid techniques, such as SMOTE Tomek Link and SMOTE ENN, performed somewhat worse than SMOTE alone. These hybrid strategies, on the other hand, have the advantage of noise reduction, making them a more trustworthy alternative in cases where data quality and noise control are essential concerns.
- Random Oversampling is much more efficient than Random undersampling techniques in this research as the number of samples in the minority class was much lower and hence to balance the remaining classes the number of instances was reduced to that smaller number of 53 eventually making it a tiny dataset and it is evident that tiny dataset is not good for accurate prediction.
- Random Forest and Decision Tree also exhibited competitive performance, highlighting their potential practical utility for accurate diabetes diagnosis.

B. *Practical Implications*

- The results demonstrate the potential practical implications of employing machine learning algorithms and resampling techniques for accurate multi-class classification of diabetes using imbalanced datasets.
- This research obtained high accuracy because the dataset was free of outliers and preprocessing and normalization were done as per requirement.
- The findings suggest that utilizing appropriate resampling techniques, such as SMOTE and Adasyn, can effectively address the class imbalance challenge and improve the reliability of the classification accuracy.
- To further improve the accuracy and performance of the multi-class classification model Larger dataset could be used which is balanced and free from outliers. Hyperparameter tuning could also help in getting better performance.

TABLE II. Comparison of accuracies with all the classifiers and resampling techniques

MLClassifiers/ Resampling technique	**LDA**	**XGB**	**RF**	**LR**	**KNN**	**GNB**	**SVC**	**DT**
Without resampling	0.904	0.992	0.976	0.920	0.936	0.960	0.924	0.988
Smote	0.900	0.988	0.980	0.916	0.944	0.94	0.924	0.984
Adasyn	0.896	0.984	0.976	0.920	0.940	0.936	0.936	0.980
Random oversampling	0.888	0.988	0.976	0.920	0.944	0.920	0.928	0.988
SVM smote	0.888	0.984	0.976	0.924	0.940	0.940	0.932	0.992
Borderline Smote	0.892	0.992	0.980	0.920	0.944	0.948	0.924	0.980
Kmeans Smote	0.904	0.992	0.976	0.936	0.948	0.944	0.936	0.984
SmoteN	0.876	0.960	0.976	0.904	0.936	0.924	0.920	0.988
Smote ENN	0.876	0.960	0.948	0.912	0.924	0.944	0.912	0.956
Smote Tomek	0.876	0.960	0.948	0.912	0.924	0.944	0.836	0.956
Random undersampling	0.844	0.960	0.936	0.876	0.796	0.908	0.836	0.964

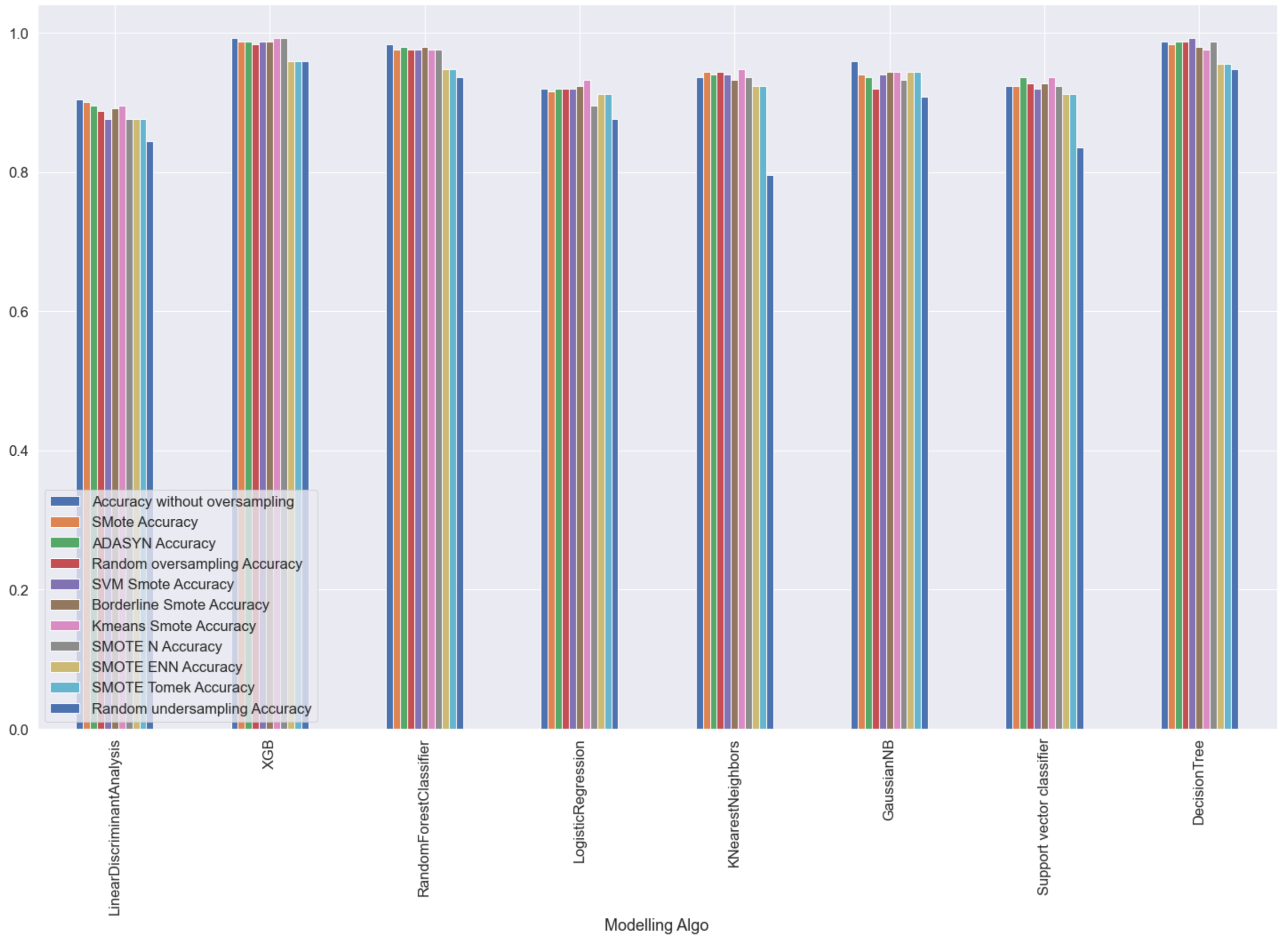

Fig. 8. Performance analysis of all the algorithms

VI. Conclusion and Future Work

- The research demonstrates the efficacy of XGBoost, Random Forest, and decision tree in accurately classifying individuals as diabetic, pre-diabetic, or non-diabetic using imbalanced datasets. Thus they are proven to be better for multiclass classification.
- Smote, SVM Smote, K means Smote and Borderline smote work comparatively well with all the classifiers.
- The results indicate the potential practical benefits of incorporating appropriate resampling techniques to improve diabetes diagnosis accuracy.
- Hybrid approaches such as Smote Tomek Lınk and Smote ENN attained good accuracy but a little lesser than SMOTE still it is more reliable because it is free from noise.
- Oversampling is much more efficient than Undersampling.
- Accuracy is higher without applying any resampling techniques because the dataset was imbalanced and it was slightly lower on applying resampling technique because the bias was reduced.

This work has been done by using the diabetes dataset but in the future other imbalanced datasets of any other disease can also be used with the proposed techniques. Using a larger dataset would be better and also the dataset used for this research is highly imbalanced, using a moderately imbalanced dataset could provide better results. Apart from accuracy, other evaluation metrics can also be taken into consideration These significant study findings contribute to a deeper understanding of the efficacy of resampling strategies and machine learning algorithms in correcting class imbalance and properly categorizing diabetes. It can help researchers, healthcare practitioners, and policymakers improve diabetes diagnosis, individualized therapy, and patient care.

References

[1] Qummar, S., Khan, F. G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z. U., Khan, I. A., & Jadoon, W. “A deep learning ensemble approach for diabetic retinopathy detection,” IEEE Access, 7, 150530-150539, 2019.

[2] ElSeddawy, A. I., Karim, F. K., Hussein, A. M., & Khafaga, D. S. “Predictive analysis of diabetes risk with class imbalance,” Computational Intelligence and Neuroscience, 2022, 3078025, 2022.

[3] Mustafa, G., Niu, Z. D., Yousif, A., & Tarus, J. “Solving the class imbalance problems using the RUSMultiBoost ensemble,” In 10th Iberian Conference on Information Systems and Technologies (pp. 1-6), 2015.

[4] Devi, D., Biswas, S. K., & Purkayastha, B. “Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance. Pattern Recognition Letters,” 93, 312, 2017.

[5] American Diabetes Association. "Diagnosis and classification of diabetes mellitus "(Position Statement). Diabetes Care, 35(Supplement 1), S64-S71, 2012.

[6] De Vegt, F., Dekker, J. M., Jager, A., et al. "Relation of impaired fasting and post-load glucose with incident type 2 diabetes in a Dutch population: the Hoorn study," Journal of the American Medical Association, 285(16), 2109-2113, 2001.

[7] The Diabetes Prevention Program (DPP) Research Group. "The diabetes prevention program (DPP): Description of lifestyle intervention" Diabetes Care, 25(12), 2165-2171, 2002.

[8] Kumar, S., Khan, M. Z., Rajendran, S., Noor, A., Dass, A. S., et al., "Imbalanced classification in diabetics using ensembled machine learning," Computers, Materials & Continua, 72(3), 4397-4409, 2022.

[9] Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J., "An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI," Sensors, 22(22), 7268, 2022.

[10] Mesquita, F., Maurício, J., & Marques, G., "Oversampling techniques for diabetes classification: A comparative study," In 2021 International Conference on e-Health and Bioengineering (EHB) (pp. 1-6), 2021.

[11] Alex, S. A., Nayahi, J. J. V., Shine, H., et al., " Deep convolutional neural network for diabetes mellitus prediction" Neural Computing and Applications, 34(4), 1319-1327, 2022.

[12] Alex, S. A., Jhanjhi, N., Humayun, M., Ibrahim, A. O., & Abulfaraj, A. W. "Deep LSTM model for diabetes prediction with class balancing by SMOTE."Electronics, 11(17), 2737, 2022.

[13] Hairani, H., Anggrawan, A., & Priyanto, D., "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link". JOIV: International Journal on Informatics Visualization, 7(1), 258-264, 2023.

[14] Shrinidhi, M., Kaushik Jegannathan, T. K., & Jeya, R. "Classification of Imbalanced Datasets Using Various Techniques along with Variants of SMOTE Oversampling and ANN" Advances in Science and Technology, 124, 504-511, 2023.

[15] Abdullah, M. N., & Wah, Y. B. "Improving diabetes mellitus prediction with MICE and SMOTE for imbalanced data." In 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 209-214), 2022.

[16] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. "Diabetes prediction using machine learning and explainable AI techniques." Healthcare Technology Letters, 10(1-2), 1-10, 2023.

[17] I. Dey and V. Pratap, "A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023, pp. 294-302