Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.11 n.51, p. 190-196. January/February., 2025.

DOI: https://doi.org/10.5935/jetia.v11i51.1453

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

INTER-CLUSTER DISTANCE-BASED SMOTE MODIFICATION FOR ENHANCED DIABETES CLASSIFICATION

Intan Nurzari[1], Ermita Sari[2], David Ibnu Harris[3], Arif Mudi Priyatno[4*] and Hidayati Rusnedy[5]

[1,2,3,4,5]Universitas Pahlawan Tuanku Tambusai, Riau, Indonesia.

[1]http://orcid.org/0009-0007-6500-1679, [2]http://orcid.org/0009-0001-2604-2693, [3]http://orcid.org/0009-0004-6834-742X, [4]http://orcid.org/0000-0003-3500-3511, [5]http://orcid.org/0009-0004-9760-4771

Email: intan.232335@universitaspahlawan.ac.id, ermita.232324@universitaspahlawan.ac.id, david.232308@universitaspahlawan.ac.id, *arifmudi@universitaspahlawan.ac.id, hidayati@universitaspahlawan.ac.id

ARTICLE INFO

Article History

Received: December 07, 2024

Revised: January 20, 2025

Accepted: January 25, 2025

Published: February 28, 2025

Keywords:

SMOTE modification,
Inter-cluster distance,
Diabetes classification,
Class imbalance,

ABSTRACT

Diabetes is a significant global health challenge, with early diagnosis playing an important role in preventing serious complications. However, medical datasets often exhibit class imbalance, where the number of non-diabetes cases is much larger than diabetes cases. This imbalance causes machine learning models to be biased towards the majority class, thus degrading prediction performance on the minority class. The problem with the commonly used oversampling method SMOTE (Synthetic Minority Oversampling Technique) is that the selection of new synthetic data formation points is done randomly, which often results in less representative synthetic data and reduces model performance. This research proposes a modification of SMOTE based on inter-cluster distance to overcome this problem. This approach uses the distance between cluster centroids in minority classes to form new synthetic data that is more representative. The research methodology involves data preprocessing, including missing value imputation, normalization, and data balancing using SMOTE modification, followed by classification using Random Forest algorithm. Evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results showed that the proposed approach achieved very high evaluation values, with accuracy, precision, recall, and F1-score of 99.7% each, far surpassing previous studies that used standard oversampling methods. This study proves that the inter-cluster distance-based SMOTE modification is effective in overcoming class imbalance and producing more representative synthetic data.

I. INTRODUCTION

Diabetes is one of the world's biggest health challenges, with a significant impact on millions of individuals each year [1]. Early diagnosis of the disease is crucial to prevent more serious complications. However, the medical data used to support diagnosis is often imbalanced [2], with the number of non-diabetic patients far outweighing those with diabetes. This imbalance causes machine learning models to be biased towards the majority class [3], reducing prediction accuracy in high-risk patients. Class imbalance is a major bottleneck in the development of reliable prediction systems for diabetes diagnosis [4].

Pima diabetes is one of the datasets often used in research to develop classification models [5]. This dataset has an uneven class distribution between diabetic and non-diabetic patients, which exacerbates the challenge of building effective prediction models. Previous studies have used various machine learning methods for diabetes detection, such as Random Forest, SVM, and artificial neural networks, with Random Forest often showing the best performance. Research by [6] performed neural network optimization for diabetes calcification. This study obtained good results above 80 percent. According to [7] did calcification with various machine learning. The results obtained are not optimal because the accuracy is less than 80 percent. This is because the research was conducted without overcoming unbalanced data.

Unbalanced data can be overcome by oversampling the minority class [8]. A commonly used method is SMOTE (Synthetic Minority Over-sampling Technique). Standard SMOTE often fails to produce realistic synthetic samples because it does not consider the local distribution or inter-cluster distance in the dataset. This

results in less representative synthetic data, which reduces the model's performance in recognizing patterns in minority classes.

Previous research has tried various approaches to overcome class imbalance in diabetes classification. Research [9] tried to apply SMOTE to overcome the imbalance. SMOTE gets accuracy values for C5.0, Random Forest, and SVM to 0.603, 0.727, and 0.727 respectively. this is because the leanring machine model occurs overfitting due to synthetic data generated by SMOTE. Research [10] conducted feature optimization and oversampling for diabetes prediction using machine learning. The results show that various SMOTE methods are used above 83 percent, best using KmeansSMote. Research [11] performed diabetes prediction using machine learning by utilizing PCA feature selection and SMOTE oversmpling. The results of increasing the minority class that has the ability to match the majority class, and the prediction results show f1-score 75 percent. Research [12] performed diabetes prediction with feature selection using Recursive Feature Elimination (RFE), and data augmentation using SMOTE (Synthetic Minority Oversampling Technique). This study achieved the highest accuracy of 82.5%, highlighting the importance of SMOTE in overcoming imbalance. Research [13] performed diabetes prediction using SMOTE and Deep learning. The results showed the highest accuracy of 86.29%, outperforming other algorithms such as Naive Bayes, Logistic Regression, and SVM. Research [14] predicting diabetes with machine learning and various oversmpling models. The results of the Multi-Layer Perceptron (MLP) Model with ADASYN resampling technique achieved the best performance, showing F1 Score 82.17 and AUC 89.61. While these approaches show promising results, they do not fully consider the importance of inter-cluster distribution in minority class datasets. Most of the previous studies relied on standard oversampling methods that only consider the distribution of the data. Previous research lacked attention to the inter-cluster distribution in minority data. This may cause the resulting synthetic data to not adequately represent the variation in the minority data.

In this study, we propose a new approach, which is a minority class distance-based SMOTE modification. By utilizing the inter-cluster distance as the basis for synthetic data generation, this approach aims to generate more representative data, improve classification accuracy, and reduce bias towards the majority class. By integrating the concept of inter-cluster distance into the synthetic data generation process, it is expected to overcome the limitations of traditional oversampling methods.

II. LITERATURE REVIEW

Previous research has identified various innovations in the application of technology to improve operational efficiency and risk management in various sectors. Research [15] used machine learning for diabetes prediction. The results show that the use of machine learning for diabetes prediction without smote can produce a classification accuracy of 80.79 percent. This research has not handled data imbalance. Research [16] performed diabetes prediction using SMOTE and ADASYN oversampling. The results showed that oversampling using ADASYN obtained accuracy, precision, recall, and f1-score results of 88.5, 82, 80, and 81 percent, respectively. This shows that the regular SMOTE method needs to be modified to improve performance in modeling. Research [17] proposed a framework for diabetes prediction that integrates oversampling techniques using SMOTE with various machine learning algorithms. The results show there is an increase in accuracy by using SMOTE and random forest compared to without using smote. SMOTE used is still standard and does not consider the inter-cluster distance in the minority class, so the improvement is not clearly visible only in the numbers behind the comma.

Research [9] tried to apply SMOTE to overcome the imbalance. SMOTE gets the accuracy value for C5.0, Random Forest, and SVM to 0.603, 0.727, and 0.727 respectively. this is because the leanring machine model occurs overfitting due to synthetic data generated by SMOTE. Research Jiang et al.(2024) conducted feature optimization and oversampling for diabetes prediction using machine learning. The results show that various SMOTE methods are used above 83 percent, best using KmeansSMote. Research [11] performed diabetes prediction using machine learning by utilizing PCA feature selection and SMOTE oversmpling. The results of increasing the minority class that has the ability to match the majority class, and the prediction results show f1-score 75 percent.

Research [12] performed diabetes prediction with feature selection using Recursive Feature Elimination (RFE), and data augmentation using SMOTE (Synthetic Minority Oversampling Technique). This study achieved the highest accuracy of 82.5%, highlighting the importance of SMOTE in overcoming imbalance. Research [14] performed diabetes prediction using SMOTE and Deep learning. The results showed the highest accuracy of 86.29%, outperforming other algorithms such as Naive Bayes, Logistic Regression, and SVM. Research [14] did diabetes prediction with machine learning and various oversmpling models. The results of the Multi-Layer Perceptron (MLP) Model with ADASYN resampling technique achieved the best performance, showing F1 Score 82.17 and AUC 89.61. The results with oversmpling are not optimal because the oversampling method technique used has not considered the inter-cluster distance in the minority data.

This research proposes a modification of SMOTE based on minority class inter-cluster distance. By utilizing the inter-cluster distance as the basis for synthetic data formation, this approach aims to generate more representative data, improve classification accuracy, and reduce bias towards the majority class. By integrating the concept of inter-cluster distance into the synthetic data generation process, it is expected to overcome the limitations of traditional oversampling methods.

III. MATERIALS AND METHODS

The main stages of this research are divided into 4 main stages. These stages are dataset, preprocessing, classification, and evaluation. Data preprocessing is done to fix missing values, outliers, normalization, and data balancing using SMOTE modification. Classification is done using random forest machine learning. Evaluation is used, namely accuracy, precision, recall, and f1-score. Figure 1 is the stages of this research.

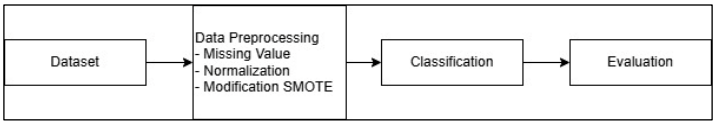

Figure 1: Stages of Research.
Source: Authors, (2025).

III.1 DATA

The data used in this study was taken from the Indian Pima Diabetes Dataset [18], which is an open dataset available in the UCI Machine Learning repository. This dataset contains medical information from 768 female individuals of Pima Indian descent who are at least 21 years old. The main focus of this dataset is to detect the presence of diabetes based on medical examination results and relevant risk factors. The dataset consists of 9 columns

that include input and output variables. The input variables include Pregnancies (number of pregnancies), Glucose (blood glucose level), BloodPressure (diastolic blood pressure in mmHg), SkinThickness (skin thickness in mm), Insulin (serum insulin level in μU/mL), BMI (Body Mass Index in kg/m²), Diabetes Pedigree Function (genetic history of diabetes), and Age (age in years). The output variable is Outcome, which is a binary indicator for diabetes, with a value of 1 indicating the individual is diagnosed with diabetes and a value of 0 indicating the individual does not have diabetes.

This dataset has zero values appearing in certain variables such as SkinThickness, Insulin, and BloodPressure. These zero values most likely reflect unrecorded or missing data, so further handling is needed. Handling was done with the median value of each class. In addition, the distribution of each variable was examined to identify potential anomalies or class imbalances in the target outcome variables. From a total of 768 samples, class imbalance was found, where 500 individuals were not diagnosed with diabetes (class 0) and 268 individuals were diagnosed with diabetes (class 1).

III.2 NORMALIZATION

Normalization is an important step in data processing that aims to equalize the scale of all features in the dataset [19]. This step is necessary so that machine learning algorithms do not give more weight to features with higher scaled values. In this research, the normalization process is performed using the Min-Max Scaling method, which is a technique that transforms feature values into the range [0,1]. This technique helps improve the stability of the model and speeds up the convergence process during training.

For example, Glucose variables that have high values tend to dominate Diabetes Pedigree Function variables that have smaller values. Without normalization, the model risks being biased towards features with larger scales. In addition, normalization also reduces the influence of extreme values or outliers, such as those found in Insulin or BMI variables, so that the model can be trained better. Thus, the application of normalization is expected to create a more stable, fair, and accurate prediction model. In this study, normalization is applied to the eight main input variables in the Indian Pima Diabetes dataset, namely Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The normalization process is performed using Equation 1.

$$\boldsymbol{x}_{\text{scaled}} = \frac{\boldsymbol{x}-\boldsymbol{x}_{\min}}{\boldsymbol{x}_{\max}-\boldsymbol{x}_{\min}} \quad (1)$$

This formula converts the original value (x) to a value in the range of 0 to 1 based on the minimum ($x_{\min}$) and maximum ($x_{\max}$) of each feature. With this method, the values of each feature can be compared fairly without being affected by the original scale. As an illustration, the Age variable, which ranges from 21 to 81 years old, is normalized to the same scale as the BMI variable, which has a value range from 18 to 67 kg/m².

III.3 SMOTE MODIFICATION

SMOTE (Synthetic Minority Oversampling Technique) is a method to deal with the problem of data imbalance in machine learning [20], specifically when there are very few minority classes in the dataset compared to the majority classes. This imbalance often results in models prioritizing the majority class and ignoring the minority class. SMOTE works by creating synthetic samples for minority classes instead of simply duplicating existing data. This technique helps reduce the possibility of overfitting that often occurs when using only simple repetition methods. The main stages of SMOTE are identifying minority classes, finding nearest neighbors, selecting neighbors for oversampling, and generating synthetic samples. Identify the minority class in the dataset, which is the class that has far fewer samples than the majority class. For the minority class, find the nearest neighbors using methods such as k-Nearest Neighbors (k-NN). The distance metric used is the Euclidean distance. From the list of found nearest neighbors, SMOTE randomly selects a number of neighbor samples to be used in the interpolation process. Interpolation to generate synthetic samples with equation 2. where $x_i$is the original data of the minority class, $x_j$is the selected neighbor, and δ is a random value between 0 and 1. This process creates a new data point that lies between the original data pair and its neighbor.

$$\boldsymbol{x_{new}} = \boldsymbol{x_i} + \boldsymbol{\delta} \times (\boldsymbol{x_j} - \boldsymbol{x_i}) \quad (2)$$

SMOTE modification is carried out at the stage of determining the location point for the formation of synthetic data. The main stages of the SMOTE modification are identification of minority classes and clustering, calculation of midpoints between clusters, determination of synthetic data locations, and generation of synthetic data. In the minority class identification and clustering stage, the clustering method used is K-means. Each cluster is represented by its centroid, which is the average point of the data in the cluster. The determination of the midpoint (M) between clusters uses Equation 3. where C1 and C2 are the centroids of the two clusters under consideration. Determination of the location of synthetic data based on the midpoint (M) using Equation 4. After the location of the synthetic point is determined, the generation of synthetic data is done with Equation 2. Algorithm 1 is the steps of Inter-Cluster Distance-Based SMOTE Modification.

$$\boldsymbol{M} = \frac{\boldsymbol{C_1}+\boldsymbol{C_2}}{\boldsymbol{2}} \quad (3)$$

$$\boldsymbol{X_{new}} = \boldsymbol{M} + \boldsymbol{\delta} \quad (4)$$

Algorithm 1: Inter-Cluster Distance-Based SMOTE Modification.

Input:	Unbalanced dataset (X,y), minority class C_{min}, number of clusters $K_{cluster}$, number of desired synthetic samples N
Output:	Dataset with extended minority class X_{new},y_{new}.
Process:	
1.	Identification of minority classes C_{min} in the dataset.
2.	Perform data grouping in C_{min} into $K_{cluster}$ clusters using an algorithm such as K-Means. Store the centroid of each cluster as C_1,C_2,....,$C_{kluster}$
3.	Calculate the midpoint between pairs of cluster centroids (C_i,C_j) for all i ≠ j : $M_{ij} = \frac{C\_i + C\_j}{2}$
4.	For each center point $\boldsymbol{M_{ij}}$, add a small variation δ to determine the new synthetic location: $\boldsymbol{x_{new}} = \boldsymbol{M_{ij}} + \boldsymbol{\delta}$ where δ is a small random value to introduce variation.
5.	Generate N new synthetic samples by repeating steps 3 and 4 until the number of synthetic samples is reached.
6.	Add synthetic samples X_{new} to the original dataset.
7.	Merge the original dataset with the synthetic dataset: X_{final} = X ∪ X_{new}, y_{final} = y ∪ y_{new}

Source: Authors, (2025).

III.4 CROSS-VALIDATION

The division of data in this study is done to ensure that the model built has good predictive ability and can be generalized to new data [21],[22]. The Indian Pima Diabetes dataset is divided

into two main subsets, namely training data and testing data. The data division process is carried out using the cross-validation method to maintain a proportional class distribution between the training and testing sets. Cross-validation was 10 folds. Cross-validation helps avoid bias that may arise due to class imbalance.

III.5 RANDOM FOREST CLASSIFICATION

Random Forest is an ensemble-based machine learning algorithm used for classification and regression tasks [23]. It is built on the principle of *bagging* (bootstrap aggregating) using decision trees as its base model. In classification, Random Forest generates predictions by combining decisions from many decision trees to improve accuracy and reducing the risk of overfitting. The main principles in random forest are:

1. Gini Index

The Gini Index measures the impurity of a node by calculating the probability of misclassification if the data is randomly selected based on the class distribution. The smaller the Gini value, the purer the node, so the feature that results in the largest Gini decrease is selected for splitting.

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{C} p_i^2} \quad (5)$$

2. Entropy

Entropy measures the disorder in the distribution of classes in a node. A low entropy value indicates that the data in the node is more homogeneous. Features with the largest entropy decrease are prioritized to splitting the data.

$$\text{Entropy} = -\sum_{i=1}^{C} \boldsymbol{p_i \log_2(p_i)} \quad (6)$$

3. Feature Importance

Feature importance indicates the relative contribution of each feature to the model's predictions. This value is calculated based on the average impurity reduction (Gini or Entropy) across all trees caused by the feature, helping to identify the most relevant features in the dataset.

III.6 EVALUATION

The evaluation metrics used include accuracy, precision, recall, and F1-score, each of which provides a different perspective on the model's performance. Accuracy measures the overall percentage of correct predictions, while precision focuses on the model's ability to identify individuals who actually have diabetes out of all those predicted to be positive. Equation 7 is how accuracy is calculated [24],[25], and Equation 8 is how precision is calculated [26],[27]. Recall assesses the model's ability to detect individuals diagnosed with diabetes out of the total diabetes cases, and F1-score provides a balance between precision and recall. Equation 9 is calculate recall [28], and Equation 10 is calculate F1-score [26]. These metrics provide an overall picture of the model's ability to provide accurate, consistent, and fair predictions.

$$\boldsymbol{Accuracy = \frac{TP+TN}{TP+TN+FP+FN}} \quad (7)$$

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \quad (8)$$

$$\boldsymbol{Recall = \frac{TP}{TP+FN}} \quad (9)$$

$$\boldsymbol{F1-Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision-Recall}} \quad (10)$$

Where TP (True Positives) is the correct prediction as positive. TN (True Negatives) is a true prediction as negative. FP (False Positives) is a false prediction as positive. FN (False Negatives) is a false prediction as negative.

IV. RESULTS AND DISCUSSIONS

This section describes the results in accordance with the research steps of Figure 1, as well as the discussion. The data used is the Indian Pima Diabetes Dataset from UCI Machine learning. The dataset contains a total of 768 samples, it is known that there is a class imbalance, where 65.1 percent are not diabetic and 34.9 percent are diabetic. The number of 65.1 percent non-diabetic individuals is 500 individuals, while the 34.9 percent who are diabetic is 268 individuals. Figure 2 is a visualization of this unbalanced (original) data distribution.

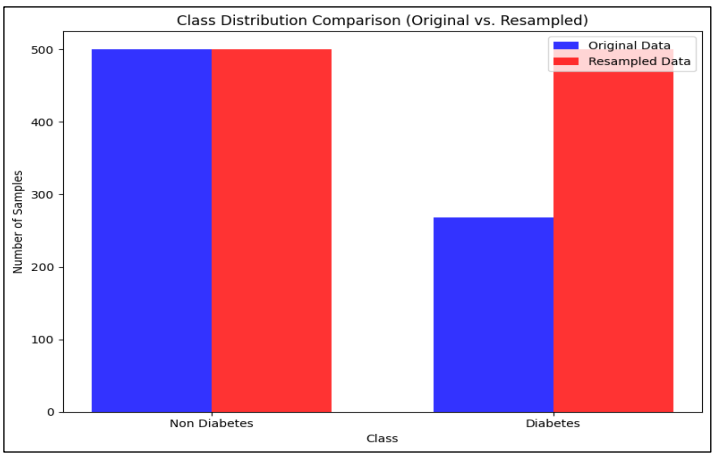

Figure 2: Class Distribution Comparision (Original vs Reampled). Source: Authors, (2025).

The original dataset has zero values in the variables Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, and BMI. This can be seen from Table 1. Pregnancies can contain zero because they have never been pregnant. While the variables Glucose, BloodPressure, SkinThickness, Insulin, and BMI cannot be zero. These zero values reflect unrecorded or missing data, so further handling is needed. Handling is done with the median value of each class. Table 2 shows the results after handling the null values.

Based on Table 2, after handling the 0 values found in some variables (which may be placeholders for missing or invalid data), the descriptive statistics for the cleaned dataset show significant changes. The handling of 0 values in the Pima Indians Diabetes dataset has resulted in a more realistic data distribution. Some variables, such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI, which previously had a value of 0, have been corrected and replaced with more reasonable values, avoiding further distortion of the analysis. These improvements provide a more accurate picture of the distribution of the variables in the dataset, enabling more effective modeling to predict diabetes in individuals based on relevant health factors.

The preprocessing data is balanced by oversampling using SMOTE modification. Figure 2 shows the result of oversampling the minority class. The results show that between the majority class and the previous minority class, the number is now the same. Data

that has been balanced is normalized with a range of 0 to 1. After normalization, modeling is then carried out using machine learning random forest classification. In the modeling process, the data is divided into 2, namely training data and test data using cross-validation. Cross-validation is used as much as 10 kfold. The test results obtained accuracy, precision, recall, and f1-score of 99.7, 99.7, 99.7, and 99.69 percent, respectively.

Table 1: Description of the indian pima diabetes dataset.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	Bmi	Diabetes Pedigree Function	Age
Count	768	768	768	768	768	768	768	768
Mean	3.85	120.9	69.12	20.54	79.8	31.99	0.47	33.24
Std	3.37	31.97	19.36	16.95	115.24	7.88	0.33	11.76
Min	0	0	0	0	0	0	0.08	21
25%	1	99	62	0	0	27.3	0.24	24
50%	3	117	72	23	30.5	32	0.37	29
75%	6	140.25	80	32	127.25	36.6	0.63	41
Max	15	199	122	99	846	67.1	2.42	81

Source: Authors, (2025).

Table 2: Description of the indian pima diabetes dataset after null value handler.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	Bmi	Diabetes Pedigree Function	Age
Count	768	768	768	768	768	768	768	768
Mean	3.85	121.67	72.39	29.09	141.75	32.43	0.47	33.24
Std	3.37	30.46	12.11	8.89	89.1	6.88	0.33	11.76
Min	0	44	24	7	14	18.2	0.08	21
25%	1	99.7	64	24	102.5	27.5	0.24	24
50%	3	117	72	28	102.5	32.05	0.37	29
75%	6	140.25	80	32	169.5	36.6	0.63	41
Max	15	199	122	99	846	67.1	2.42	81

Source: Authors, (2025).

Table 3. Comparison of evaluation results with previous research.

Reference	Accuracy	Precision	Recall	F1-Score
[15]	80.79	73.68	75	-
[16]	88.5	82	80	81
[17]	89.6	86	84	85.2
[9]	72.7	-	-	-
[10]	88.56	-	-	86.66
[11]	-	89	65	75
[12]	82.5	-	-	-
[13]	86.29	81.9	84.2	-
[14]	-	-	-	82.18
Proposed	**99.7**	**99.7**	**99.7**	**99.69**

Source: Authors, (2025)

A comparison of the evaluation results in Table 3 shows the significant achievements of our proposed method, compared to previous studies. These studies used various machine learning techniques and oversampling methods to overcome class imbalance in diabetes prediction. Research [15] achieved an accuracy of 80.79% with a machine learning approach without using SMOTE, which shows the limitations of an imbalanced dataset. Research [16],[17] integrated oversampling techniques such as SMOTE and ADASYN, resulting in accuracies of 88.5% and 89.6%, respectively, with moderate improvements in precision, recall, and F1-score metrics. Research [10] used KMeans-SMOTE, achieving 88.56% accuracy with further feature optimization. Reseach [11],[12] utilized PCA and Recursive Feature Elimination (RFE) for feature selection, combined with SMOTE, resulting in F1-score of 75% and 82.5% respectively. However, the performance is still below that of the proposed method. Deep learning approaches also show potential, such as by [13] who achieved 86.29% accuracy using SMOTE and deep convolutional neural network, while [14] reported an F1-score of 82.18% with ADASYN and MLP models.

Our proposed method achieves significantly superior performance on all metrics, with accuracy, precision, recall, and F1-score of 99.7% each. By modifying SMOTE using inter-cluster distance analysis, the method effectively addresses class imbalance while preserving the underlying data structure, ensuring balanced learning between classes. This achievement confirms the importance of developing oversampling techniques and integrating a robust classification framework for diabetes prediction. The proposed methodology not only overcomes class imbalance, but also achieves unprecedented prediction performance, setting a new standard for research in this field.

V. CONCLUSIONS

This study proposes a modification of the inter-cluster distance-based SMOTE method to address data imbalance in

diabetes prediction using the Indian Pima Diabetes dataset. This approach is designed to generate more representative synthetic data by considering the inter-cluster distribution of minority classes, thus improving the quality of the classification model. The results showed that the proposed method achieved significantly higher evaluation performance than previous studies, with accuracy, precision, recall, and F1-score of 99.7% each. Moreover, this modification significantly reduces the bias towards the majority class while preserving the underlying data structure. Based on these results, it can be concluded that the proposed method successfully improves classification accuracy, and reduces bias towards the minority class. This success opens up opportunities for further application in other disease diagnosis, especially on datasets with high class imbalance.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Intan Nurzari, Ermita Sari, and David Ibnu Harris.
Methodology: Intan Nurzari, Arif Mudi Priyatno and Hidayati Rusnedy.
Investigation: Hidayati Rusnedy.
Discussion of results: Intan Nurzari, Ermita Sari, David Ibnu Harris, Arif Mudi Priyatno and Hidayati Rusnedy.
Writing – Original Draft: Intan Nurzari, Ermita Sari, David Ibnu Harris
Writing – Review and Editing: Intan Nurzari, Arif Mudi Priyatno and Hidayati Rusnedy.
Resources: David Ibnu Harris.
Supervision: Arif Mudi Priyatno and Hidayati Rusnedy
Approval of the final text: Intan Nurzari, Ermita Sari, David Ibnu Harris, Arif Mudi Priyatno and Hidayati Rusnedy

VIII. REFERENCES

[1] M. Gollapalli *et al.*, "A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM," *Comput. Biol. Med.*, vol. 147, p. 105757, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105757.

[2] R. Vij and S. Arora, "A novel deep transfer learning based computerized diagnostic Systems for Multi-class imbalanced diabetic retinopathy severity classification," *Multimed. Tools Appl.*, vol. 82, no. 22, pp. 34847–34884, Sep. 2023, doi: 10.1007/s11042-023-14963-4.

[3] P. Sampath *et al.*, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Sci. Rep.*, vol. 14, no. 1, p. 28984, Nov. 2024, doi: 10.1038/s41598-024-78519-8.

[4] K. Ahnaf Alavee *et al.*, "Enhancing Early Detection of Diabetic Retinopathy Through the Integration of Deep Learning Models and Explainable Artificial Intelligence," *IEEE Access*, vol. 12, pp. 73950–73969, 2024, doi: 10.1109/ACCESS.2024.3405570.

[5] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, p. e24536, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24536.

[6] A. F. Ashour, M. M. Fouda, Z. M. Fadlullah, and M. I. Ibrahem, "Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, IEEE, Apr. 2024, pp. 1–7. doi: 10.1109/ICMI60790.2024.10585703.

[7] S. Jain, S. K. Sunori, A. Mittal, and P. Juneja, "Detection of Diabetes using Various Machine Learning Techniques," in *8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2024 - Proceedings*, IEEE, Oct. 2024, pp. 1382–1386. doi: 10.1109/I-SMAC61858.2024.10714839.

[8] A. I. ElSeddawy, F. K. Karim, A. M. Hussein, and D. S. Khafaga, "Predictive Analysis of Diabetes-Risk with Class Imbalance," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–16, Oct. 2022, doi: 10.1155/2022/3078025.

[9] M. Khairul Rezki, M. I. Mazdadi, F. Indriani, M. Muliadi, T. H. Saragih, and V. A. Athavale, "Application Of SMOTE To Address Class Imbalance In Diabetes Disease Classification Utilizing C5.0, Random Forest, And SVM," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 343–354, Aug. 2024, doi: 10.35882/jeeemi.v6i4.434.

[10] L. Jiang *et al.*, "A feature optimization study based on a diabetes risk questionnaire," *Front. Public Heal.*, vol. 12, no. 1, Feb. 2024, doi: 10.3389/fpubh.2024.1328353.

[11] S. R. Velu, V. Ravi, and K. Tabianan, "Machine learning implementation to predict type-2 diabetes mellitus based on lifestyle behaviour pattern using HBA1C status," *Health Technol. (Berl).*, vol. 13, no. 3, pp. 437–447, Jun. 2023, doi: 10.1007/s12553-023-00751-5.

[12] E. Sabitha and M. Durgadevi, "Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 921–930, 2022, doi: 10.14569/IJACSA.2022.01309107.

[13] S. A. Alex, J. J. V. Nayahi, H. Shine, and V. Gopirekha, "Deep convolutional neural network for diabetes mellitus prediction," *Neural Comput. Appl.*, vol. 34, no. 2, pp. 1319–1327, Jan. 2022, doi: 10.1007/s00521-021-06431-7.

[14] M. Talebi Moghaddam *et al.*, "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 220, Sep. 2024, doi: 10.1186/s12874-024-02341-z.

[15] A. Pyne and B. Chakraborty, "Artificial Neural Network based approach to Diabetes Prediction using Pima Indians Diabetes Dataset," in *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*, IEEE, May 2023, pp. 01–06. doi: 10.1109/ICCAD57653.2023.10152382.

[16] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.

[17] L. Al-dabbas, "Early Detection of Female Type-2 Diabetes using Machine Learning and Oversampling Techniques," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1237–1245, Sep. 2024, doi: 10.47738/jads.v5i3.298.

[18] M. Kahn, "Diabetes - UCI Machine Learning Repository."

[19] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 24, no. 1, p. 337, Sep. 2023, doi: 10.1186/s12859-023-05465-z.

[20] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, "Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 36, Dec. 2022, doi: 10.1186/s12911-022-01775-z.

[21] M. Bhagat and B. Bakariya, "Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach," *Natl. Acad. Sci. Lett.*, vol. 45, no. 5, pp. 401–404, Oct. 2022, doi: 10.1007/s40009-022-01131-9.

[22] A. M. Priyatno, W. F. Ramadhan Sudirman, and R. J. Musridho, "Feature selection using non-parametric correlations and important features on recursive feature elimination for stock price prediction," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 2, p. 1906, Apr. 2024, doi: 10.11591/ijece.v14i2.pp1906-1915.

[23] U. e Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," *Sensors*, vol. 22, no. 14, p. 5247, Jul. 2022, doi: 10.3390/s22145247.

[24] A. M. Priyatno and L. Ningsih, "TF - IDF Weighting to Detect Spammer Accounts on Twitter based on Tweets and Retweet Representation of Tweets," *Sist. J. Sist. Inf.*, vol. 11, no. 3, pp. 614–622, 2022, [Online]. Available: http://sistemasi.ftik.unisi.ac.id/index.php/stmsi/issue/view/46

[25] A. M. Priyatno, "SPAMMER DETECTION BASED ON ACCOUNT, TWEET, AND COMMUNITY ACTIVITY ON TWITTER," *J. Ilmu Komput. dan Inf.*, vol. 13, no. 2, pp. 97–107, Jul. 2020, doi: 10.21609/jiki.v13i2.871.

[26] A. M. Priyatno and F. I. Firmananda, "N-Gram Feature for Comparison of Machine Learning Methods on Sentiment in Financial News Headlines," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 1, no. 1, pp. 01–06, Jul. 2022, doi: 10.31004/riggs.v1i1.4.

[27] A. M. Priyatno, M. M. Muttaqi, F. Syuhada, and A. Z. Arifin, "Deteksi Bot Spammer Twitter Berbasis Time Interval Entropy dan Global Vectors for Word Representations Tweet's Hashtag," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 5, no. 1, p. 37, Jan. 2019, doi: 10.26594/register.v5i1.1382.

[28] M. R. A. Prasetya and A. M. Priyatno, "Dice Similarity and TF-IDF for New Student Admissions Chatbot," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 1, no. 1, pp. 13–18, Jul. 2022, doi: 10.31004/riggs.v1i1.5.