A Modified SMOTE with Noise Filtering and Manhattan Distance Metric Approach to Address Imbalanced Health Datasets

Triyanna Widiyaningtyas
Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia
triyannaw.ft@um.ac.id (corresponding author)

Hairani Hairani
Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia | Department of Computer Science, Universitas Bumigora, Mataram, Indonesia
hairani.2305349@students.um.ac.id

Didik Dwi Prasetya
Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia
didikdwi@um.ac.id

Utomo Pujianto
Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia
utomo.pujianto.ft@um.ac.id

Wahyu Caesarendra
Department of Mechanical and Mechatronics Engineering, Faculty of Engineering and Science, Curtin University Malaysia, Malaysia
w.caesarendra@curtin.edu.my

Received: 5 May 2025 | Revised: 26 May 2025, 9 June 2025, 15 June 2025, and 18 June 2025 | Accepted: 21 June 2025

 DOI: https://doi.org/10.48084/etasr.11925

ABSTRACT

Imbalanced class distribution remains a significant challenge in healthcare data analysis, particularly in disease-related datasets where minority classes representing critical conditions such as diabetes are severely underrepresented. This disproportionate representation often results in biased predictive models that exhibit reduced sensitivity to minority classes, leading to suboptimal diagnostic accuracy and reduced generalizability. Imbalanced data can decrease the performance of classification methods and result in overfitting. SMOTE is a frequently used method for addressing data imbalance. A recent SMOTE variant considers only outliers to remove minority classes (data noise) without considering minority data neighboring majority classes, which are considered noise. This research aimed to modify SMOTE based on KNN filtering and a modification of Manhattan-based distance metrics to reduce the generation of noise data in minority classes and minimize overlap. The proposed method is called NR-Modified SMOTE and has several stages in balancing data: (i) filtering by removing minority classes close to majority classes (data noise) using the KNN method, and (ii) applying SMOTE oversampling with the modification of the Manhattan distance metric. Experiments were carried out on two health datasets, Pima and Haberman, with NR-Modified SMOTE and classification using Random Forest, SVM, and Naive Bayes using 10-fold cross-validation, where the proposed method led to better accuracy for all classifiers than NR-SMOTE without distance metric modifications.

Keywords-SMOTE modification; distance metric modification; filtering approach; noise reduction

I. INTRODUCTION

Health data often encounters imbalanced data issues, more frequently compared to other fields, particularly in disease-related datasets such as breast cancer [1], diabetes, stroke [2], and heart disease [3]. Imbalanced health data is a crucial challenge to address. Imbalanced data refers to an unequal distribution of classes within a dataset, where one class is more dominant than the others [4]. Imbalanced data leads to biased classification outcomes [5], overfitting, and poor performance [6] of classification methods. SMOTE is a baseline solution to address imbalanced data issues, generating synthetic data based on linear interpolation between samples of the minority class with K Nearest Neighbors (KNN) [7].

However, the SMOTE method has limitations, such as generating noisy synthetic data that can lead to overfitting [8, 9] and overlapping between classes in decision boundary areas. Several studies have attempted to address the shortcomings of the SMOTE approach. The SMOTE-LOF method [10] combines SMOTE with the Local Outlier Factor (LOF) algorithm to retain samples from the original minority class and remove minority data considered outliers. In [11], the Radius-SMOTE approach was proposed to address data imbalance issues, focusing on noise data filtering and generating synthetic data based on a safe radius to prevent overlap between classes. In [12], Borderline-SMOTE was modified with noise reduction techniques on uneven data. In [13], a K-Means approach was combined with SMOTE (KM-SMOTE). In [14-16], synthetic instances for minority categories were generated using different distance metrics in SMOTE, observing improvements in classification performance depending on the metric used. In [17], the Outlier-SMOTE approach was proposed to address imbalanced COVID-19 data, focusing on creating minority classes in outlier areas. In [18], KSMOTE was introduced, which applied a Kalman filter to identify and remove noisy data samples, reducing the impact of overfitting on classification methods. In [19], an Adaptive Synthetic (Adasyn) approach introduced modifications to distance metrics, using the Manhattan distance to resolve imbalanced student graduation data.

The baseline reference for this study to modify the SMOTE method is [10]. The LOF technique is typically used in outlier identification, but it can be used to detect noise because most outliers are considered noise. SMOTE-LOF focuses on detecting noise among minority instances located in outlier areas. Minority-class instances identified as outliers are removed while retaining samples of the original minority class. Based on the findings of this study, the SMOTE-LOF method with k=3 performed better than the unmodified SMOTE on Pima, Haberman, and Glass datasets. However, this study has several shortcomings that must be addressed. First, only minority data considered noise in outlier areas are addressed, while important minority noise data adjacent to majority classes should also be resolved. Second, the study uses health data as experimental material, so removing minority classes from outlier areas may not be appropriate, as they may contain important information. Therefore, it is necessary to involve SMOTE oversampling processes, as proposed in [17], which suggests the Outlier-SMOTE oversampling technique to improve COVID-19 detection by prioritizing minority class instances in outlier areas. Third, the study in [10] does not address the problem of overlapping instances of minority classes generated by SMOTE, which can potentially become noise. An approach that could be used to minimize the overlap produced by SMOTE is to modify distance metrics, as in [14-16].

This study adopts a combination of filtering and distance metric modification strategies to address the issues of noise and overlapping that commonly arise from SMOTE. The filtering mechanism is applied before SMOTE to eliminate minority samples that are likely to be noise, contributing positively to the handling of imbalanced data, as supported in [20, 21]. Additionally, modifying the distance metric within SMOTE helps reduce the overlap between synthetic minority samples and instances of the majority class, which can otherwise introduce more noise [22]. Therefore, this study aimed to modify SMOTE based on KNN filtering and Manhattan-based distance metrics to reduce noise data in minority classes and minimize overlapping.

In this context, this study proposes a novel method termed Noise Reduction (NR)-Modified SMOTE (NR-Modified SMOTE), to improve the original SMOTE by incorporating two key innovations: (i) a KNN-based filtering process is used to discard minority samples located near majority instances, as they are more likely to be noise, and (ii) the Manhattan distance metric is utilized in the SMOTE process to reduce data overlap in synthetic minority class generation. Although the use of filtering techniques and Manhattan distance metrics in SMOTE has been widely discussed, the novelty in this study lies in the integration and application sequence of these components, which have not been examined in previous studies. Specifically, this study proposes a filtering algorithm for noise reduction prior to the synthetic sample generation process, followed by a modification of the Manhattan distance metric in SMOTE to maintain the distribution structure of the minority class. As this combination and application sequence have not been explicitly explored previously, this study offers a more effective approach to maintaining interclass boundaries and reducing the overlap of synthetic data with the majority class. The main contributions of this study include:

- Integrates a KNN-based filtering approach to eliminate noisy minority class instances close to the majority class.
- Introduces Manhattan distance into the SMOTE framework to reduce the occurrence of overlapping synthetic data, thus improving class separability.

II. RESEARCH METHOD

The first step is a review of the literature to collect and analyze relevant studies on modifications of the SMOTE method to address data problems. Literature review analysis yields insights into the latest approaches to modifying SMOTE for imbalanced data resolution and identifies gaps from previous research. The second step involves collecting health

data from Kaggle, including two publicly available datasets: the Pima dataset [23], which contains diagnostic data to predict the onset of diabetes, and the Haberman dataset [24], which includes survival records of breast cancer patients. The Pima dataset comprises 768 data entries with eight features, while the Haberman dataset includes 306 entries and three features. These two datasets were chosen because they are the most frequently used in experiments related to cases of imbalanced data [25]. Table I shows details of these two datasets.

TABLE I. DATASET USED

Dataset	Examples	Minority	Majority
Pima	768	268	500
Haberman	306	81	225

The data preprocessing steps used in this study include scaling and data balancing using the proposed method. Feature scaling is applied to reduce the disproportionate influence of attributes with larger value ranges over those with smaller ranges. The scaling process is carried out using:

$$X' = \frac{x - min_x}{max_x - min_x} \quad (1)$$

SMOTE, introduced in [26], is an oversampling technique to address class imbalance by generating synthetic minority instances through linear interpolation with Euclidean distance. This process, described in (2) and (3) [27], identifies the k nearest neighbors of the minority samples and interpolates among them. However, Euclidean distance may be less effective in high-dimensional spaces, potentially introducing noise and overlap in synthetic data. As an alternative, the Manhattan distance (4) has been proposed [28], offering improved performance in high-dimensional contexts when used within the SMOTE framework [29].

$$y' = y^i + (y^j - y^i) * \Upsilon \quad (2)$$

where term y' refers to the inclusion of the minority class, y^i denotes the minority class instance, y^j represents the value of the k-nearest neighbor to y^i, and Υ is a randomly selected vector with a value within the range of 0 to 1.

$$ED(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)} \quad (3)$$

$$D_{Manhattan}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad (4)$$

where ED refers to the Euclidean distance, $D_{Manhattan}$ denotes the Manhattan distance, x_i represents the i-th test data on variable x, and y_i represents the i-th sample data on variable y.

SMOTE has a significant drawback, as it may generate noisy data within the synthesized minority classes, and overlapping data may introduce additional noise, which can result in inaccurate classification outcomes. To address this issue, NR-Modified SMOTE utilizes a filtering approach with the KNN method to remove noise from minority classes adjacent to majority classes, with k=3 as in [11, 30, 31], and then performs SMOTE with a modification of the Manhattan distance metric to reduce potentially noise-generating overlapping data.

In [10], the focus was solely on resolving the noise of minority classes located in outlier areas while neglecting the noise of minority classes adjacent to the majority classes and not addressing the potentially noise-generating overlapping data produced by SMOTE. In contrast, the proposed solution involves resolving noise from minority classes adjacent to majority classes using a filtering approach and addressing potentially noise-generating overlapping data by modifying the Manhattan distance metric in SMOTE. Figure 1 shows the detailed process of the proposed method. The initial step focuses on filtering through the KNN method to classify minority data as either noise or non-noise categories [31]. Minority data instances that are close to the majority class (noise) are removed. Then, the SMOTE method modified with the Manhattan distance metric is employed to balance the minority class and minimize the potential overlap of artificially generated minority class data, which could result in noise. By removing noise from the minority class before applying SMOTE, the amount of noise generated is reduced, thereby improving the overall performance of the classification method [32]. The fourth step consists of splitting the data into training and testing sets through a 10-fold cross-validation. This technique partitions the data into 10 subsets, with each fold alternating between serving as training and testing data. The next step involves applying classification methods, specifically Random Forest (RF), SVM, and Naïve Bayes (NB).

The last stage involves assessing the effectiveness of the classification method. This evaluation is carried out using a confusion matrix, which illustrates the counts of correct and incorrect classifications, as presented in Table II [33, 34]. Accuracy is used to evaluate the performance of the models [35-38].

TABLE II. CONFUSION MATRIX

Actual	Prediction	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

III. RESULTS AND DISCUSSION

A. Experimental Design

The Pima dataset consists of 768 samples, with 500 belonging to the Negative class and 268 to the Positive class, resulting in an imbalance ratio of 1.87. The Haberman dataset includes 306 samples, with 225 from the positive class and 81 from the negative class, leading to an imbalance ratio of 2.78.

Subsequently, a scaling procedure is applied to reduce the influence of features with the largest value range (max) compared to those with the smallest value range (min). Following this, data balancing is performed on both the Pima and Haberman datasets using the proposed NR-Modified SMOTE method. Figure 2 shows the sampling process using NR-Modified SMOTE on the Pima and Haberman datasets.

The dataset, balanced using the proposed method, is subsequently classified using the RF, SVM, and NB algorithms using 10-fold cross-validation. Classification performance is evaluated using accuracy, which is derived from the confusion matrix.

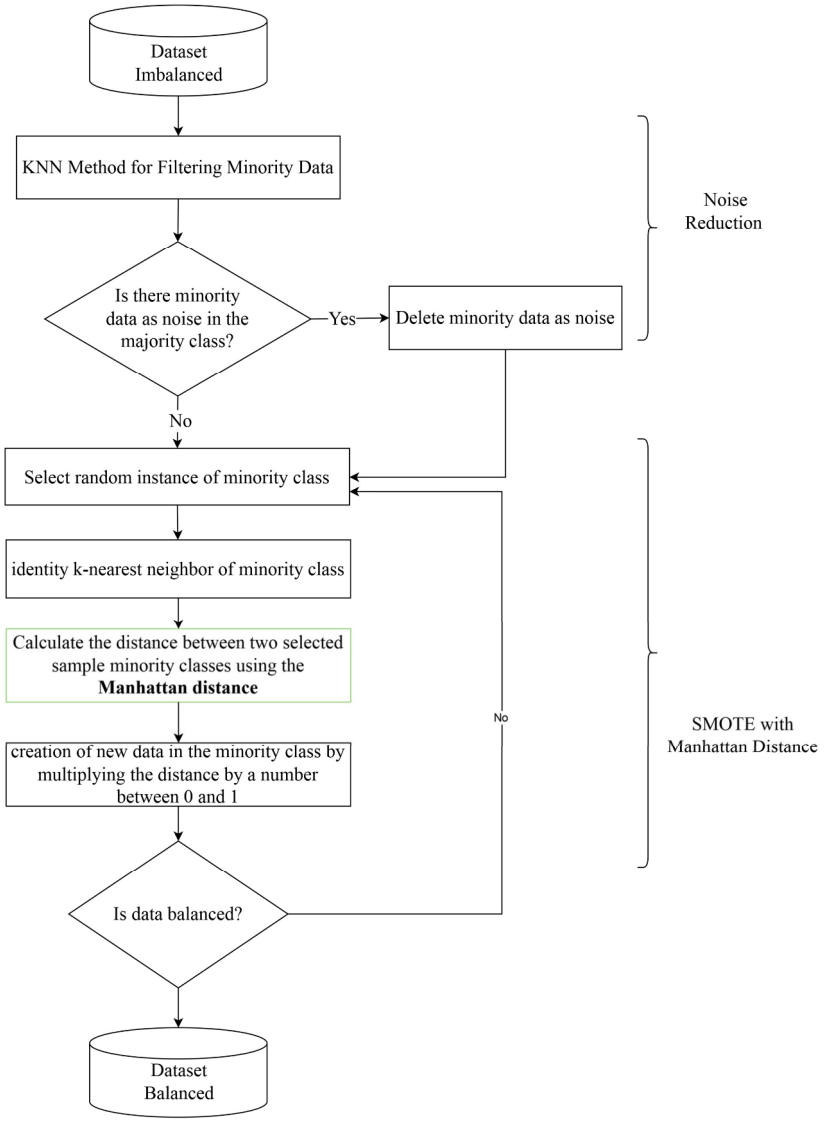

Fig. 1. Proposed method: NR-Modified SMOTE.

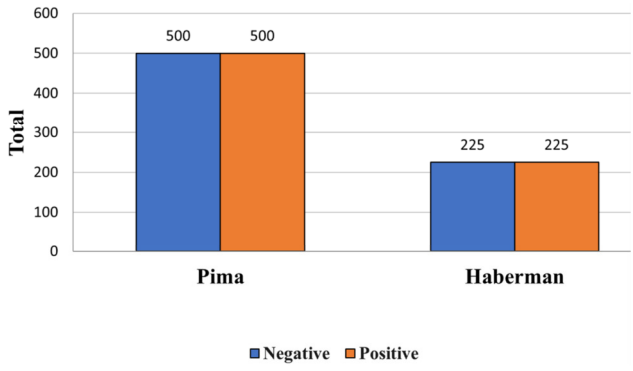

Fig. 2. Number of cases after oversampling with NR-Modified SMOTE.

The original data distribution, shown in Figure 3, shows a significant class imbalance, characterized by the dominance of the number of samples in the majority class and a limited distribution in the minority class, which tends to be concentrated in certain feature areas and accompanied by the emergence of outliers in extreme areas. After applying the NR-Modified SMOTE method (Figure 4), the class distribution becomes more balanced by adding synthetic data to the minority class that is more evenly distributed in the feature space. This technique avoids areas containing noise, thereby reducing the potential for overlapping between classes.

B. Experimental Result

Tables III and IV present the confusion matrix results of the RF method with NR-Modified SMOTE on the Pima and the Haberman datasets, respectively. Tables V and VI show the confusion matrix results of the SVM method with NR-Modified SMOTE on the Pima and the Haberman datasets, respectively. Finally, Tables VII and VIII show the confusion matrix results of the NB method with NR-Modified SMOTE on the Pima and the Haberman datasets, respectively.

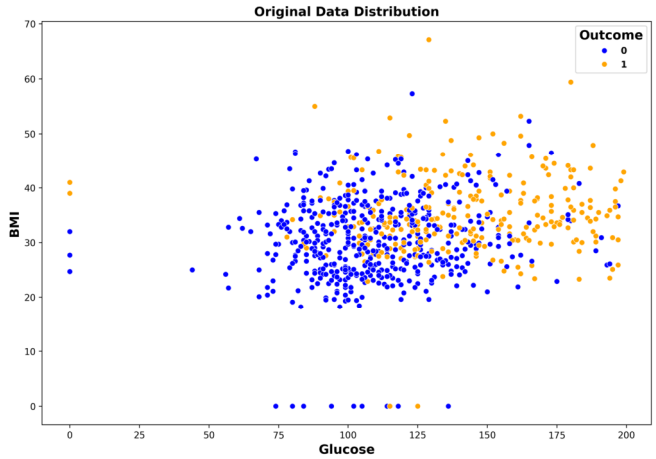

Fig. 3. Original data distribution.

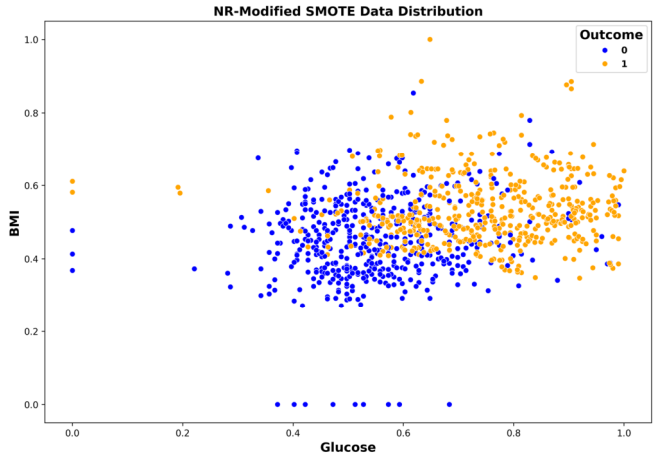

Fig. 4. Data distribution after applying NR-Modified SMOTE.

TABLE III. CONFUSION MATRIX FOR PIMA DATA USING RF

Actual	NR-SMOTE		NR-Modified SMOTE	
	Negative	Positive	Negative	Positive
Negative	425	75	428	73
Positive	35	466	34	466

TABLE IV. CONFUSION MATRIX FOR HABERMAN DATA WITH RF

Actual	NR-SMOTE		NR-Modified SMOTE	
	1	2	1	2
1	189	39	185	40
2	9	216	5	220

TABLE V. CONFUSION MATRIX FOR PIMA DATA WITH SVM

Actual	NR-SMOTE		NR-Modified SMOTE	
	Negative	Positive	Negative	Positive
Negative	401	99	400	100
Positive	70	430	63	437

TABLE VI. CONFUSION MATRIX FOR HABERMAN DATA WITH SVM

Actual	NR-SMOTE		NR-Modified SMOTE	
	1	2	1	2
1	188	37	190	35
2	76	149	75	150

TABLE VII. CONFUSION MATRIX RESULTS FOR PIMA DATA WITH NB

Actual	NR-SMOTE		NR-Modified SMOTE	
	Negative	Positive	Negative	Positive
Negative	403	97	403	97
Positive	100	400	105	395

TABLE VIII. CONFUSION MATRIX ON HABERMAN DATA WITH SVM

Actual	NR-SMOTE		NR-Modified SMOTE	
	1	2	1	2
1	198	27	195	30
2	92	133	88	137

According to these tables, the NR-Modified SMOTE method led to high accuracy across both the Pima and Haberman datasets with all classifiers. In the Pima dataset, RF correctly classified 428 out of 500 negative and 466 out of 500 positive instances (Table III), while in the Haberman dataset, RF classified 185 out of 225 for class 1 and 220 out of 225 for class 2 (Table IV). In the Pima dataset, SVM classified 400 out of 500 negative and 437 out of 500 positive instances (Table V), while in the Haberman dataset, SVM classified 190 out of 225 for class 1 and 150 out of 225 for class 2 (Table VI). In the Pima dataset, NB classified 403 out of 500 negative and 397 out of 500 positive instances (Table VII), while in the Haberman dataset, Naïve Bayes classified 195 out of 225 for class 1 and 137 out of 225 for class 2 (Table VIII).

Table IX compares the performance of NR-SMOTE and NR-Modified SMOTE with RF, SVM, and NB on the Pima and Haberman datasets. NR-Modified SMOTE outperforms NR-SMOTE in all cases. With RF, NR-Modified SMOTE achieves 89.36% accuracy on Pima and 89.20% accuracy on Haberman. Similarly, SVM with NR-Modified SMOTE also shows 83.70% accuracy on Pima and 75.64% accuracy on Haberman. Overall, NR-Modified SMOTE with RF outperforms SVM and NB on both datasets.

TABLE IX. PERFORMANCE RESULTS OF THE PROPOSED METHOD WITH DIFFERENT CLASSIFIERS ON HEALTH DATASETS

Distance	Method	Dataset	Accuracy
Euclidean	Random Forest	Pima	89.03%
	SVM		83.10%
	Naïve Bayes		80.30%
	Random Forest	Haberman	88.78%
	SVM		74.89%
	Naïve Bayes		73.56%
Manhattan	Random Forest	Pima	89.36%
	SVM		83.70%
	Naïve Bayes		79.80%
	Random Forest	Haberman	89.20%
	SVM		75.64%
	Naïve Bayes		73.78%

C. Discussion

The test results indicate that the RF method with NR-Modified SMOTE exhibits superior performance compared to the SVM and NB methods across all the datasets utilized. Moreover, NR-Modified SMOTE demonstrates better accuracy compared to NR-SMOTE without distance metric modification in all classification methods employed. The NR-Modified SMOTE method shows superior performance compared to the traditional SMOTE method in dealing with data imbalance problems. This is due to the initial stage in the form of a noise reduction process that aims to eliminate minority data that are not representative or can cause overlapping between classes. After the noise reduction process, synthetic data is added using SMOTE with the Manhattan distance approach, which is more effective in minimizing the risk of overlapping between classes compared to using Euclidean distance.

Figure 5 presents a comparative visualization of the test results of the SMOTE, NR-SMOTE, and NR-Modified SMOTE methods on the two benchmark datasets, using the same classification algorithm, namely RF, to maintain the validity of the comparison. The results show that NR-Modified SMOTE provides the best performance. On the Pima dataset, the accuracy increases from 82.22% (SMOTE) and 89.03% (NR-SMOTE) to 89.36%. Meanwhile, on the Haberman dataset, the accuracy increases from 66.89% (SMOTE) and 88.78% (NR-SMOTE) to 89.20%. These findings indicate that the NR-Modified SMOTE method is more effective in addressing class imbalance, especially on datasets with small sizes and uneven class distributions.

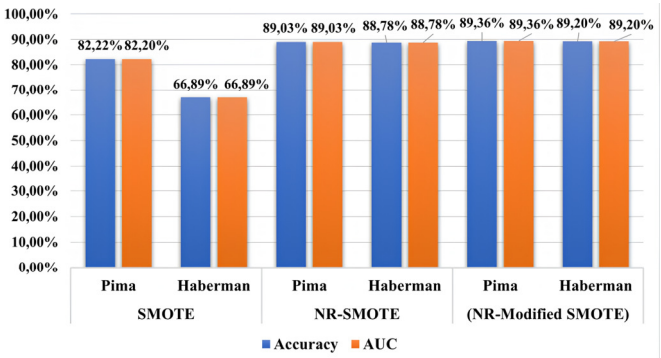

Fig. 5. Performance comparison of SMOTE, NR-SMOTE, and NR-Modified SMOTE using RF.

When compared to previous studies, the proposed NR-Modified SMOTE method demonstrates enhanced performance compared to SMOTE-LOF [10], Radius-SMOTE [11], and IRS-BAG-Integrated Radius-SMOTE [38] on the same dataset. As shown in Table X, NR-Modified SMOTE improves accuracy by 15.14% on the Pima dataset and 20.32% on the Haberman dataset over SMOTE-LOF [10]. Compared to Radius-SMOTE [11], accuracy is increased by 2.96% on the Pima dataset and 12.6% on the Haberman dataset. Compared to IRS-BAG-Integrated Radius-Smote [38], accuracy is increased by 8.36% on the Pima dataset and 8.20% on the Haberman dataset.

TABLE X. COMPARISON OF THE PROPOSED WITH PREVIOUS APPROACHES

Method	Dataset	Accuracy	AUC
SMOTE-LOF [10]	Pima	74.22%	74.24%
	Haberman	68.88%	-
Radius-SMOTE [11]	Pima	86.40%	-
	Haberman	76.60%	-
IRS-BAG-Integrated Radius-Smote [38]	Pima	81.00%	86.00%
	Haberman	81.00%	79.00%
Adasyn with RF	Pima	81.11%	81.22%
	Haberman	60.30%	60.10%
SMOTE with RF	Pima	82.22%	82.20%
	Haberman	66.89%	66.89%
NR-SMOTE with RF	Pima	89.03%	89.03%
	Haberman	88.78%	88.78%
Proposed method (NR-Modified SMOTE)	Pima	89.36%	89.36%
	Haberman	89.20%	89.20%

The proposed approach demonstrates superior performance across all classification techniques compared to various SMOTE variants, including SMOTE-LOF [10], Radius-SMOTE [11], and IRS-BAG-Integrated Radius-SMOTE [38]. The improved performance of NR-Modified SMOTE is attributed to the noise removal process and the adjustment of the Manhattan distance metric within SMOTE. By removing noise from the minority class before applying SMOTE, the noise introduced by SMOTE itself is reduced, thus increasing the accuracy of classification [32, 39, 40]. Furthermore, using SMOTE with a modified Manhattan distance to balance minority classes helps reduce the likelihood of noise arising from artificial overlaps in the minority class [19]. As shown in Table X, NR-Modified SMOTE surpasses NR-SMOTE, which lacks distance metric adjustment, in both datasets, aligning with the findings from previous studies [16, 19].

The utilization of the Manhattan distance is superior to the Euclidean distance for generating synthetic minority classes because of its ability to capture nonlinear relationships and handle high-dimensional data more effectively. However, the impact of Manhattan distance in minimizing overlapping data in synthetic minority classes is not significantly pronounced. This could be influenced by the absence of decision boundaries in generating synthetic minority classes [41], thus necessitating the creation of new decision boundaries using clustering approaches to focus more on generating minority classes within each cluster. This approach could reduce data overlapping more effectively, improving the performance of classification methods. Another limitation of this study is that it uses only two small-scale datasets and does not use high-dimensional datasets such as MIMIC-III.

IV. CONCLUSION

This study proposes the NR-Modified SMOTE approach with RF to address data imbalance in healthcare datasets such as Pima and Haberman. The proposed method outperforms several previous studies on the same data. This is attributed to removing minority data deemed noise near the majority class using KNN filtering, followed by the Manhattan distance in SMOTE to reduce overlapping data in synthetic minority classes, potentially considered noise. The proposed method performs better than SVM and NB, with accuracies of 89.36% and 89.20% on the Pima and Haberman datasets, respectively.

In addition, NR-Modified SMOTE shows a better performance improvement compared to SMOTE-LOF, Radius-SMOTE, and IRS-BAG-Integrated Radius-SMOTE on the Pima and Haberman datasets. Compared to SMOTE-LOF, accuracy increases by 15.14% on the Pima dataset and bt 20.32% on the Haberman dataset. Compared to Radius-SMOTE, accuracy increases by 2.96% on the Pima dataset and by 12.6% on the Haberman dataset. Compared to IRS-BAG-Integrated Radius-Smote, accuracy increases by 8.36% on the Pima dataset and by 8.20% on the Haberman dataset. Further research could improve this method by implementing clustering processes to group data into several clusters to establish decision boundaries. Subsequently, SMOTE could be applied to each cluster to reduce overlapping data from synthetic minority classes. However, a limitation of this study is that it uses only two small-scale datasets and does not use high-dimensional datasets such as MIMIC-III.

ACKNOWLEDGMENT

This research was supported by Universitas Negeri Malang in an Expertise Group (KBK) research scheme with contract number 24.2.198/UN32.14.1/LT/2025

REFERENCES

[1] L. G. R. Putra, K. Marzuki, and H. Hairani, "Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction," *Engineering and Applied Science Research*, vol. 50, 2023, Art. no. 577583, https://doi.org/10.14456/EASR.2023.59.

[2] Q. Yin, X. Ye, B. Huang, L. Qin, X. Ye, and J. Wang, "Stroke Risk Prediction: Comparing Different Sampling Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, https://doi.org/10.14569/IJACSA.2023.01406115.

[3] K. Wang *et al.*, "Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using SMOTE+ENN and Machine Learning," *Risk Management and Healthcare Policy*, vol. Volume 14, pp. 2453–2463, Jun. 2021, https://doi.org/10.2147/RMHP.S310295.

[4] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, Jun. 2024, Art. no. 122778, https://doi.org/10.1016/j.eswa.2023.122778.

[5] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201, Aug. 2019, https://doi.org/10.29207/resti.v3i2.945.

[6] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, https://doi.org/10.1109/ACCESS.2021.3074243.

[7] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, https://doi.org/10.1007/s10994-022-06296-4.

[8] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing*, vol. 143, Aug. 2023, Art. no. 110415, https://doi.org/10.1016/j.asoc.2023.110415.

[9] V. W. de Vargas, J. A. S. Aranda, R. Dos Santos Costa, P. R. Da Silva Pereira, and J. L. V. Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowledge and Information Systems*, vol. 65, no. 1, pp. 31–57, Jan. 2023, https://doi.org/10.1007/s10115-022-01772-8.

[10] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3413–3423, Jun. 2022, https://doi.org/10.1016/j.jksuci.2021.01.014.

[11] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, https://doi.org/10.1109/ACCESS.2021.3080316.

[12] M. Revathi and D. Ramyachitra, "A Modified Borderline Smote with Noise Reduction in Imbalanced Datasets," *Wireless Personal Communications*, vol. 121, no. 3, pp. 1659–1680, Dec. 2021, https://doi.org/10.1007/s11277-021-08690-y.

[13] Q. Liu *et al.*, "Application of KM-SMOTE for rockburst intelligent prediction," *Tunnelling and Underground Space Technology*, vol. 138, Aug. 2023, Art. no. 105180, https://doi.org/10.1016/j.tust.2023.105180.

[14] Q. Dai, J. Liu, and J. L. Zhao, "Distance-based arranging oversampling technique for imbalanced data," *Neural Computing and Applications*, vol. 35, no. 2, pp. 1323–1342, Jan. 2023, https://doi.org/10.1007/s00521-022-07828-8.

[15] S. Feng, J. Keung, P. Zhang, Y. Xiao, and M. Zhang, "The impact of the distance metric and measure on SMOTE-based techniques in software defect prediction," *Information and Software Technology*, vol. 142, Feb. 2022, Art. no. 106742, https://doi.org/10.1016/j.infsof.2021.106742.

[16] A. Balakrishnan, J. Medikonda, P. K. Namboothiri, and M. Natarajan, "Mahalanobis Metric-based Oversampling Technique for Parkinson's Disease Severity Assessment using Spatiotemporal Gait Parameters," *Biomedical Signal Processing and Control*, vol. 86, Sep. 2023, Art. no. 105057, https://doi.org/10.1016/j.bspc.2023.105057.

[17] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intelligence-Based Medicine*, vol. 3–4, Dec. 2020, Art. no. 100023, https://doi.org/10.1016/j.ibmed.2020.100023.

[18] G. S. Thegas, Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets," *Machine Learning with Applications*, vol. 8, Jun. 2022, Art. no. 100267, https://doi.org/10.1016/j.mlwa.2022.100267.

[19] H. A. Gameng, B. D. Gerardo, and R. P. Medina, "A Modified Adaptive Synthetic SMOTE Approach in Graduation Success Rate Classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 3053–3057, Dec. 2019, https://doi.org/10.30534/ijatcse/2019/63862019.

[20] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, "SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling," *Applied Intelligence*, vol. 51, no. 3, pp. 1394–1409, Mar. 2021, https://doi.org/10.1007/s10489-020-01852-8.

[21] Q. Chen, Z. L. Zhang, W. P. Huang, J. Wu, and X. G. Luo, "PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets," *Neurocomputing*, vol. 498, pp. 75–88, Aug. 2022, https://doi.org/10.1016/j.neucom.2022.05.017.

[22] E. Elyan, C. F. Moreno-Garcia, and C. Jayne, "CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification," *Neural Computing and Applications*, vol. 33, no. 7, pp. 2839–2851, Apr. 2021, https://doi.org/10.1007/s00521-020-05130-z.

[23] "Pima Indians Diabetes Database." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[24] "Haberman's Survival Data Set." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set.

[25] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1310–1318, Sep. 2024, https://doi.org/10.62527/joiv.8.3.2283.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, https://doi.org/10.1613/jair.953.

[27] E. Blanco-Mallo, L. Morán-Fernández, B. Remeseiro, and V. Bolón-Canedo, "Do all roads lead to Rome? Studying distance measures in the context of machine learning," *Pattern Recognition*, vol. 141, Sep. 2023, Art. no. 109646, https://doi.org/10.1016/j.patcog.2023.109646.

[28] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Applied Soft Computing*, vol. 76, pp. 380–389, Mar. 2019, https://doi.org/10.1016/j.asoc.2018.12.024.

[29] X. Gao and G. Li, "A KNN Model Based on Manhattan Distance to Identify the SNARE Proteins," *IEEE Access*, vol. 8, pp. 112922–112931, 2020, https://doi.org/10.1109/ACCESS.2020.3003086.

[30] G. M. Lin and H. C. Zeng, "Electrocardiographic Machine Learning to Predict Mitral Valve Prolapse in Young Adults," *IEEE Access*, vol. 9, pp. 103132–103140, 2021, https://doi.org/10.1109/ACCESS.2021.3098039.

[31] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 563–597, Jun. 2016, https://doi.org/10.1007/s10844-015-0368-1.

[32] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5059–5074, Sep. 2022, https://doi.org/10.1016/j.jksuci.2022.06.005.

[33] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 2, pp. 343–348, Jun. 2022, https://doi.org/10.30630/joiv.6.2.985.

[34] Sucipto, D. D. Prasetya, and T. Widiyaningtyas, "A Supervised Hybrid Weighting Scheme for Bloom's Taxonomy Questions using Category Space Density-based Weighting," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 22102–22108, Apr. 2025, https://doi.org/10.48084/etasr.10226.

[35] H. Qteat, M. Awad, and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 11–22, Jun. 2021, https://doi.org/10.22266/ijies2021.0630.02.

[36] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, "An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning," *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 2, pp. 306–316, Jun. 2022, https://doi.org/10.30630/joiv.6.2.990.

[37] Z. Farou, M. Aharrat, and T. Horváth, "A Comparative Study of Assessment Metrics for Imbalanced Learning," in *New Trends in Database and Information Systems*, vol. 1850, A. Abelló, P. Vassiliadis, O. Romero, R. Wrembel, F. Bugiotti, J. Gamper, G. Vargas Solar, and E. Zumpano, Eds. Springer Nature Switzerland, 2023, pp. 119–129.

[38] L. Yuningsih, G. A. Pradipta, D. Hermawan, P. D. W. Ayu, D. P. Hostiadi, and R. R. Huizen, "IRS-BAG-Integrated Radius-SMOTE Algorithm with Bagging Ensemble Learning Model for Imbalanced Data Set Classification," *Emerging Science Journal*, vol. 7, no. 5, pp. 1501–1516, Oct. 2023, https://doi.org/10.28991/ESJ-2023-07-05-04.

[39] N. A. Firdausanti, I. Mendonça, and M. Aritsugi, "Noise-free sampling with majority framework for an imbalanced classification problem," *Knowledge and Information Systems*, vol. 66, no. 7, pp. 4011–4042, Jul. 2024, https://doi.org/10.1007/s10115-024-02079-6.

[40] D. D. Prasetya, T. Widiyaningtyas, H. Hairani, and A. Aminuddin, "Addressing Imbalance in Health Datasets: A New Method NR-Clustering SMOTE and Distance Metric Modification," *Computers, Materials & Continua*, vol. 82, no. 2, pp. 2931–2949, 2025, https://doi.org/10.32604/cmc.2024.060837.

[41] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Information Sciences*, vol. 572, pp. 574–589, Sep. 2021, https://doi.org/10.1016/j.ins.2021.02.056.