Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 1, pp. 130-140, February 2026

RESEARCH PAPER

OPEN ACCES

The Effect of Smote-Tomek on the Classification of Chronic Diseases Based on Health and Lifestyle Data

Muhammad Adika Riswanda, Friska Abadi, Muhammad Itqan Mazdadi, Mohammad Reza Faisal, and Rudy Herteno

Department of Computer Science, Lambung Mangkurat University, Kalimantan Selatan, Indonesia

Abstract

Machine learning models for chronic disease prediction are often trained on imbalanced healthcare datasets, where non-disease cases dominate. This condition can lead to misleadingly high accuracy while failing to identify patients with chronic diseases, limiting clinical usefulness. This study aims to analyze the impact of class imbalance on model performance and to evaluate the effectiveness of the SMOTE–Tomek resampling technique in improving chronic disease prediction. This research provides empirical evidence that accuracy alone is insufficient for evaluating healthcare models and demonstrates that imbalance-aware preprocessing is essential for valid and reliable chronic disease detection. Five classification models, such as Support Vector Machine, Random Forest, K-Nearest Neighbors, Gradient Boosting, and XGBoost, were evaluated on a lifestyle-based chronic disease dataset under two conditions: without resampling and with SMOTE–Tomek. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC. Without SMOTE–Tomek, all models failed to detect chronic disease cases, producing near-zero recall and F1-scores despite accuracy exceeding 80%. After applying SMOTE–Tomek, substantial improvements were observed across all models, particularly in recall and AUC. Support Vector Machine achieved the best overall performance, with an accuracy of 92.9%, a precision of 92%, a recall of 93.9%, an F1-score of 0.93, and an AUC of 0.98. The findings confirm that handling class imbalance is a prerequisite for meaningful chronic disease prediction. The consistent increase in recall and AUC across all evaluated models confirms that the improvement stems from enhanced class separability rather than metric inflation. The proposed approach supports more reliable early screening and decision-support systems in preventive healthcare.

Paper History

Received Dec. 10, 2025
Revised Feb. 15, 2026
Accepted Feb. 25, 2026
Published March 3, 2026

Keywords

Imbalance;
SMOTE-Tomek;
Chronic Disease;
Machine Learning;
Healthcare Analytics

Author Email

2011016210025@mhs.ulm.ac.id
friska.abadi@ulm.ac.id
mazdadi@ulm.ac.id
reza.faisal@ulm.ac.id
rudy.herteno@ulm.ac.id

I. Introduction

Chronic diseases are a crucial issue in health and a serious concern globally. These long-term medical conditions, such as cardiovascular disease, cancer, kidney disease, and diabetes, not only cause high morbidity and mortality rates but also place a significant economic burden on society [1], [2]. The increase in the prevalence of chronic diseases, driven by an aging population and lifestyle changes, underscores the urgency of developing innovative, accurate prediction models for early detection and personalized intervention. Early detection plays an important role in preventing disease progression, reducing related complications, and improving the overall prognosis of affected individuals. Identifying diseases at an early stage is crucial to minimizing severity, especially considering that some types of chronic diseases, such as heart disease, diabetes, and certain cancers, are often asymptomatic in the early stages, so delayed detection can lead to significant and irreversible progression [3].

By definition, chronic diseases are health conditions or disorders that last for at least 3 months and can cause serious long-term effects [4]. These diseases are more common in older adults and can generally be controlled, but are difficult to cure. Common forms of chronic diseases include cancer, cardiovascular disease (CVD), diabetes, brain disease, liver disease, stroke, and arthritis [5]. The WHO reports that chronic diseases cause 41 million deaths each year, or about 74% of all global deaths. A total of 17 million deaths occur in individuals under the age of 70, with only 15% of these premature deaths occurring in high-income countries. CVD is the leading cause of death, followed by cancer and diabetes, while the main risk factors include smoking, lack of physical activity, excessive alcohol consumption, and poor diet [6].

In the field of health informatics, chronic disease prediction plays a very important role. Chronic disease diagnosis systems can assist in designing appropriate treatments and improving the quality of patient care [7]. Early detection is the only effective way to reduce

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

mortality rates and prepare for future disease management, so that patients can receive appropriate care and the severity of the disease can be prevented. However, there are certainly challenges in developing early detection in health informatics, such as health data imbalance.

Data imbalance in a dataset can be addressed using a data sampling approach that aims to balance the number of samples across classes. Several methods commonly used to address this problem include oversampling, undersampling, and hybrid sampling. The oversampling method adds samples to the minority class to make the data distribution more balanced [8], while undersampling reduces the number of samples in the majority class to achieve a more balanced data distribution [9]. In addition, the hybrid sampling approach combines oversampling and undersampling techniques simultaneously to improve the quality of unbalanced datasets [10]. One technique often used in this category is SMOTE-Tomek Link, which combines the Synthetic Minority Over-sampling Technique (SMOTE) with the Tomek Link strategy to reduce data redundancy and improve classification model accuracy [11]. Compared to pure oversampling methods such as SMOTE or ADASYN, which may increase class overlap by generating synthetic samples in noisy regions, and pure undersampling methods that risk discarding informative majority instances, SMOTE–Tomek provides a more balanced mechanism by simultaneously synthesizing minority samples and removing borderline majority samples. This dual strategy is particularly suitable for healthcare datasets, where minimizing class overlap and improving clarity of decision boundaries are essential for reliable chronic disease prediction.

SVMs are effective supervised learning model for classification and regression, especially in high-dimensional spaces with clear margin separation through optimal hyperplanes [12]. Unlike SVMs, Random Forest are ensemble method that builds multiple decision trees to improve accuracy and reduce the risk of overfitting [13]. Meanwhile, Gradient Boosting combines weak models incrementally to form a stronger predictive model, with each iteration improving on previous errors [14]. XGBoost is a gradient boosting library optimized for efficiency and flexibility. This algorithm excels in classification and regression, handles missing values, and applies regularization to prevent overfitting, making it effective for large-scale data [15]. K-Nearest Neighbors (KNN) is a non-parametric regression method that uses the nearest data points in a dataset to estimate new data values [16]. By constructing a state vector from current and past data, this algorithm computes the shortest distance between the initial and current vectors to identify the nearest neighbors [17].

Recent studies have explored diverse machine learning approaches for chronic disease prediction, emphasizing feature selection strategies, ensemble learning, and algorithm optimization. Ghosh et al. [18] reported high predictive accuracy on Relief-based Random Forest, while Kumar and Sikamani [19] demonstrated the superiority of SVMs on certain chronic illness datasets. Saqlain et al. [20] and Mohan et al. [21] highlighted the effectiveness of hybrid feature-selection frameworks, and Yang et al. [22] showed that XGBoost combined with feature importance analysis can significantly improve renal risk prediction using large-scale Electronic Health Records. Collectively, these studies confirm the potential of machine learning in chronic disease analytics.

However, despite these advances, several critical gaps remain. First, many prior studies primarily focus on feature selection or model comparison, while the impact of hybrid class-imbalance handling, particularly SMOTE–Tomek has not been systematically evaluated across multiple algorithmic paradigms within a unified experimental framework. Second, although class imbalance is widely acknowledged as a major challenge in medical prediction tasks, existing research often applies a single resampling technique without thoroughly examining how hybrid oversampling–undersampling mechanisms influence model generalization and decision boundary formation. Third, limited work has investigated imbalance mitigation specifically in datasets that integrate demographic, physiological, and lifestyle-related variables simultaneously, which are increasingly relevant for preventive healthcare modeling.

To address these gaps, this study proposes a comprehensive classification framework that integrates SMOTE–Tomek hybrid resampling with five machine learning models: SVM, Random Forest, KNN, LSTM, and XGBoost. This allows for a structured comparison of imbalance effects across fundamentally different learning paradigms. Unlike previous studies that emphasize model accuracy alone, this research systematically evaluates minority-class sensitivity, discriminative capability (AUC), and overall generalization performance under both imbalanced and balanced conditions. The main contributions are: (1) establishing an imbalance-aware classification pipeline tailored to chronic disease prediction using health and lifestyle data; (2) providing a comparative multi-model assessment under identical preprocessing and resampling settings to ensure methodological consistency; (3) empirically demonstrating how hybrid resampling reshapes predictive behavior across different algorithms; and (4) offering a reproducible experimental workflow to support future development in preventive health informatics.

II. Materials and Method

This research aims to investigate the impact of class-imbalance handling on chronic disease classification by comparing the effects of applying and omitting the SMOTE–Tomek resampling strategy within the proposed machine-learning pipeline. The workflow begins with dataset acquisition and preprocessing to ensure data consistency, followed by two parallel experimental settings: one utilizing the original imbalanced dataset and the other incorporating SMOTE–Tomek to address the skewed class distribution. In the SMOTE–Tomek setting, synthetic minority-class samples are generated while overlapping majority-class instances are removed, leading to a more balanced and informative training set, whereas the non-resampled setting preserves the original data distribution as a baseline. Both datasets are

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

subsequently divided into training and testing subsets using an 80:20 ratio, after which multiple machine learning models—including Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost—are trained and optimized through hyperparameter tuning. The comparative evaluation of model performance under these two conditions enables a clear assessment of how SMOTE–Tomek influences predictive accuracy and classification stability in chronic disease detection, which can be seen in the following Fig 1:

A. Dataset

The dataset used in this study is the Synthetic Health and Lifestyle Dataset, consisting of 7,500 synthetically generated records designed to emulate realistic population-level health and lifestyle characteristics. Although artificially constructed, the dataset was created using probabilistic distributions and domain-informed assumptions derived from commonly reported epidemiological statistics and public health literature, such as typical age distributions, BMI ranges, lifestyle prevalence patterns, and chronic disease risk tendencies. These assumptions were designed to approximate realistic variable interactions and prevalence rates observed in population-level studies, thereby enhancing the dataset's representativeness for methodological evaluation purposes.

All entries are fully anonymized and contain no real human information, as the dataset is entirely synthetically

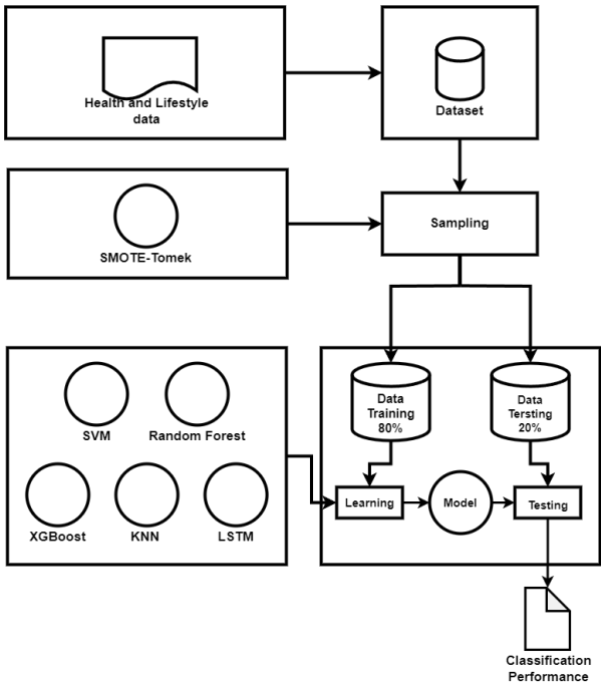

Fig. 1. Research Flowchart

generated. Consequently, no ethical approval or institutional review board (IRB) clearance was required for this study. Since the data do not originate from real individuals, there are no direct privacy risks or personally identifiable information involved. Nevertheless, the study maintains responsible data-handling practices and transparently reports the dataset's synthetic nature to avoid misinterpretations regarding its real-world clinical applicability.

The dataset includes a comprehensive set of demographic, physiological, lifestyle, and health-related variables that may serve as predictors of chronic disease. A detailed description of each feature is presented in Table 1.

Table 1. Feature Description of the Synthetic Health and Lifestyle Dataset

No	Column	Description
1	ID	Unique identifier for each individual
2	Age	Age of the individual (years)
3	Gender	Gender identity (Male, Female, Other)
4	Height_cm	Height in centimeters
5	Weight_kg	Weight in kilograms
6	BMI	Body Mass Index calculated as weight (kg) / height (m²)
7	Smoker	Indicates smoking status (Yes/No)
8	Exercise_Freq	Frequency of physical exercise (None, 1–2 times/week, 3–5 times/week, Daily)
9	Diet_Quality	Self-rated dietary quality (Poor, Average, Good, Excellent)
10	Alcohol_Consumption	Level of alcohol intake (None, Low, Moderate, High)
11	Chronic_Disease	Presence of chronic illness (Yes/No)
12	Stress_Level	Self-reported stress level on a 1–10 scale
13	Sleep_Hours	Average hours of sleep per night

B. SMOTE-Tomek

The Synthetic Health and Lifestyle Dataset exhibits a noticeable class imbalance in the target variable Chronic_Disease, where instances representing individuals without chronic disease significantly outnumber those with chronic conditions. Such an imbalance can bias the learning process toward the majority class, leading to suboptimal detection of minority-class patterns and reduced predictive reliability. To address this issue, data resampling techniques are commonly employed to rebalance class distributions and improve model generalization. Among existing sampling strategies, Synthetic Minority Over-sampling Technique (SMOTE) is widely adopted because it generates artificial minority-class samples through interpolation between nearest neighbors, rather than simple duplication, thereby preserving the underlying feature space structure.

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

However, SMOTE alone may introduce noise in regions where minority and majority classes overlap, potentially affecting the clarity of decision boundaries [23]. To mitigate this limitation, the SMOTE–Tomek method integrates SMOTE with Tomek Links, an undersampling technique that removes majority-class instances that form ambiguous nearest-neighbor pairs with minority samples, thereby enhancing class separability [24], [25].

In this study, SMOTE–Tomek is applied after data preprocessing to correct the imbalance observed in the Synthetic Health and Lifestyle Dataset. Prior to resampling, the dataset shows a skewed class distribution, with 6,052 non-chronic cases and 1,448 chronic disease cases. After applying SMOTE–Tomek, the class distribution is perfectly balanced, yielding 4,789 samples per class. This combined oversampling and undersampling process first synthesizes new chronic disease instances based on local neighborhood structures and subsequently eliminates overlapping majority-class samples, yielding a cleaner and more discriminative feature space. As a result, the balanced dataset provides a more robust foundation for downstream feature selection, as shown in Fig. 2.

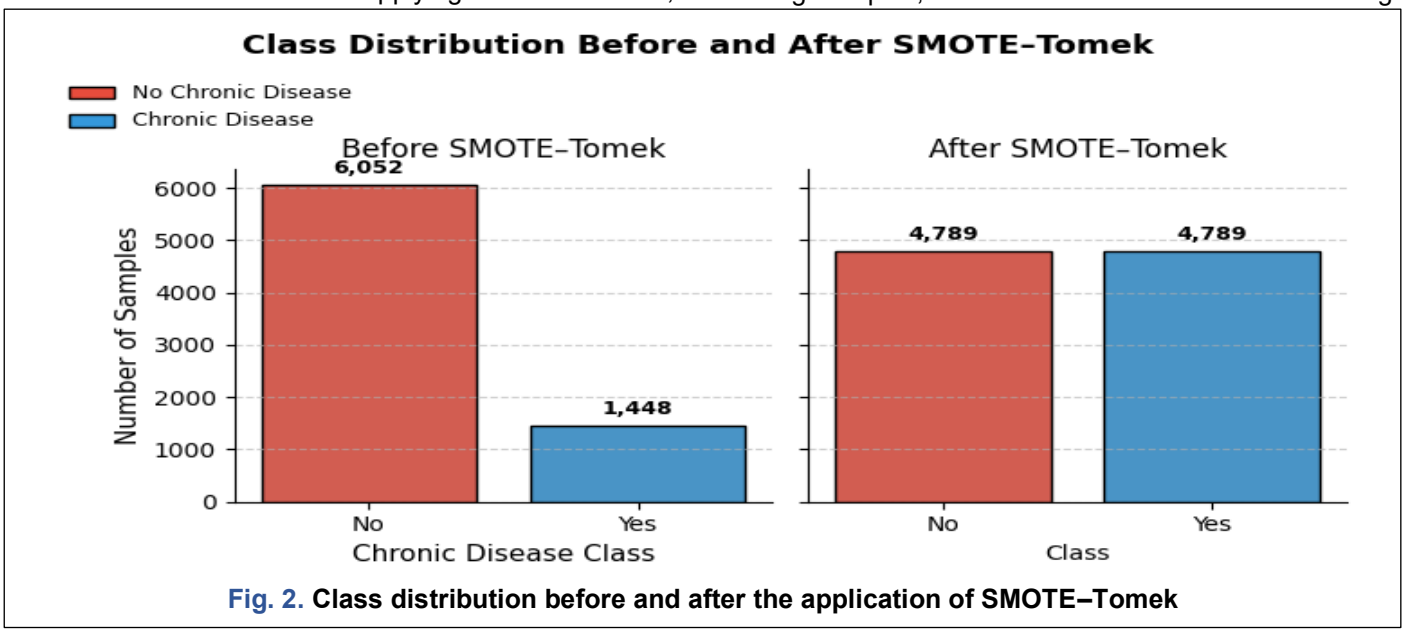

Fig. 2. Class distribution before and after the application of SMOTE–Tomek

C. Data Splitting

A widely adopted method for validating models is data splitting, which involves dividing a dataset into two subsets: training and testing. Models are trained on the training data and validated using the test data. This separation ensures that model evaluation is unbiased, allowing for an accurate assessment of predictive performance while minimizing concerns about overfitting to the training data [26]. Once a splitting ratio is determined, the previously mentioned data-splitting methods can be applied. A typical ratio is 80:20, with 80% of the data allocated to training and 20% to testing. However, other ratios like 70:30, 60:40, and even 50:50 are also commonly used in practice. No definitive rule exists for determining the optimal ratio for a particular dataset. In this research, we adopt an 80:20 data-splitting ratio, where 80% of the dataset is used for training and the remaining 20% is reserved for testing.

D. Machine Learning Model

1) Support Vector Machine (SVM)

Support Vector Machine (SVM) is widely recognized as a robust supervised learning technique applicable to both classification and regression problems. Rather than merely fitting decision boundaries, SVM seeks to determine an optimal separating hyperplane by maximizing the margin between support vectors of different classes. This margin maximization principle enables SVM to achieve strong generalization performance, particularly in scenarios involving limited training samples, where conventional statistical learning approaches often struggle to ensure optimal or stable solutions [27]. In cases where the data cannot be effectively separated in the original input space, SVM addresses this limitation by using kernel-based transformations. By mapping the input data into a higher-dimensional feature space, kernel functions facilitate clearer separation between classes that are otherwise non-linearly distributed. Commonly used kernels, including the Radial Basis Function (RBF) and polynomial kernels, have demonstrated notable effectiveness in improving classification accuracy, especially for datasets characterized by small sample sizes. The selection of an appropriate kernel function plays a critical role in accurately defining the decision boundary and reducing misclassification. Consequently, SVM performance is highly influenced by kernel characteristics such as type, scale, and surface smoothness, all of which directly affect the sharpness and reliability of class separation [28].

2) Random Forest (RF)

In tree-based learning models, the feature space is progressively partitioned into smaller, more homogeneous regions through a process known as recursive binary partitioning. This mechanism divides the input space using a sequence of decision rules defined over individual features. Each split can be mathematically

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

expressed as a threshold-based rule, as shown in Eq. (1) [29]:

$$x_j \geq a_k \quad (1)$$

Where x_j denotes the j-th input feature and a_k represents the corresponding split threshold. The partitioning procedure is applied recursively until a stopping criterion, such as maximum tree depth or minimum sample size, is reached. The resulting terminal regions, commonly referred to as leaf nodes, assign a class label or prediction value to all samples falling within that region. Through successive axis-aligned splits, decision trees are able to approximate complex, nonlinear decision boundaries by combining multiple simple partitions of the feature space [29]. Extending this framework, the Random Forest algorithm constructs an ensemble of decision trees and integrates their individual predictions to obtain a final output. Each tree independently produces a class prediction, and the ensemble decision is determined through a majority voting strategy. This aggregation process can be mathematically represented as shown in Eq. (2) [30]:

$$\hat{y} = arg\,max \sum_{t=1}^{T} \Pi(h_t(x) = c) \quad (2)$$

where $h_t(x)$ denotes the prediction of the t-th tree for input sample x, T is the total number of trees, and II(·) is an indicator function. By combining predictions from multiple decorrelated trees, Random Forest effectively reduces variance and alleviates overfitting, resulting in a more stable and generalizable classifier compared to a single decision tree model [30].

3) Xgboost

The learning mechanism of Extreme Gradient Boosting is formulated as an optimization problem that jointly considers prediction accuracy and model complexity. During the boosting process, a sequence of functions f_k is learned to iteratively correct the residual errors produced by previous trees. The predicted output for the i-th instance generated by the k-th tree is denoted as $f_k(x_i)$ and the overall objective function minimized at each iteration can be mathematically expressed as shown in (3) [31]:

$$Obj = \sum_{i=1}^{n} |(\hat{y}_i, y_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (3)$$

where the first term represents the loss function quantifying the discrepancy between the true target value y_i and its prediction $\hat{y}_i$. The second term corresponds to the regularization component, which penalizes model complexity and is defined as in (4) [31]:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \omega^2 \quad (4)$$

with T denoting the number of leaf nodes in the tree, ω representing the vector of leaf weights, and γ and λ serving as regularization parameters that control overfitting by constraining tree structure and weight magnitude [31].

Building upon this formulation, Extreme Gradient Boosting (XGBoost), proposed by Chen et al. [32], extends the conventional Gradient Boosting Decision Tree (GBDT) framework into a highly scalable and efficient tree-boosting system. By leveraging an ensemble of decision trees optimized via gradient-based learning, XGBoost can address both classification and regression problems with high predictive performance. Its design emphasizes computational efficiency and robustness, making it particularly effective for large-scale and high-dimensional datasets. The integration of second-order gradient information and explicit regularization allows XGBoost to achieve superior generalization compared to traditional boosting approaches [32].

4) KNN

Distance measurement plays a central role in the K-Nearest Neighbors (KNN) algorithm, as it determines how similarity between data points is quantified. In this study, four commonly adopted distance metrics are employed, namely Euclidean, Manhattan, Minkowski, and Chebyshev distances, each capturing different geometric interpretations of proximity within the feature space [33]. The Euclidean distance, which represents the straight-line distance between two points in an n-dimensional space, is mathematically defined as shown in (5) [34]:

$$d = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad (5)$$

where x_i and y_i denote the corresponding components of two data points [34]. Alternatively, the Manhattan distance, often referred to as the city-block distance, computes similarity by summing the absolute differences across dimensions, as expressed in (6) [35]:

$$d = \sum_{i=1}^{n} |x_i - y_i| \quad (6)$$

Which is particularly effective in grid-based or high-dimensional representations [35]. Generalizing both metrics, the Minkowski distance introduces a tunable parameter that unifies Euclidean and Manhattan distances as special cases, and can be formulated as shown in (7) [36]:

$$d = (\sum_{i=1}^{n} |x_i - y_i|^c)^{1/c} \quad (7)$$

Providing greater flexibility in adapting to different data distributions [36]. In contrast, the Chebyshev distance focuses on the maximum absolute coordinate difference between two vectors, as presented in (8) [37]:

$$d = max_i |x_i - y_u| \quad (8)$$

and is particularly suitable when the largest deviation along any dimension dominates similarity assessment [37].

Building on these distance formulations, K-Nearest Neighbors (KNN), originally introduced by Cover and Hart as a classification technique, has evolved into a highly adaptable non-parametric method for regression over the past several decades. In regression settings, KNN estimates the output of a new observation by identifying the k most similar data points in the training set and aggregating their target values, typically through averaging [38]. The algorithm constructs a state vector using current and historical observations, after which distances between the query instance and existing samples are computed to determine the nearest neighbors [33]. Once the k closest neighbors are selected, the predicted value is obtained by averaging

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

their corresponding outputs at the subsequent time step. The overall effectiveness of KNN in regression tasks is therefore strongly influenced by the choice of distance metric, as it directly governs neighbor selection and, consequently, prediction accuracy.

5) Gradient Boosting

Gradient Boosting is a predictive modeling approach that incrementally constructs an ensemble of decision trees, sequentially adding models that correct the residuals (errors) produced by their predecessors. In this framework, each decision tree focuses on learning the discrepancy between the actual values and the predictions generated by the previous model, enabling the overall model to be progressively refined by capturing information that was not adequately explained in earlier iterations [39]. By leveraging historical data that reflect demand trends, seasonal variations, and other relevant factors, Gradient Boosting can uncover complex patterns and nonlinear relationships within the data, thereby producing more accurate forecasts. Furthermore, its inherent flexibility in modeling nonlinear interactions among variables makes it particularly well-suited for environments characterized by dynamic, continuously evolving demand patterns. Consequently, the application of Gradient Boosting in inventory management has been shown to enhance forecasting accuracy, reduce excess stock levels, and improve overall operational efficiency [40].

III. Result

A. Model Performance Without SMOTE–Tomek

Prior to applying any resampling strategy, the dataset exhibits pronounced class imbalance, with non-chronic disease cases substantially outnumbering chronic disease cases. Under these conditions, all evaluated classification models were trained and tested directly on the imbalanced data to assess their baseline performance. The quantitative performance of each model without SMOTE–Tomek is summarized in Table 2.

Table 2. Performance of Classification Models Without SMOTE–Tomek

Model	Acc	Preci	Recall	F1-Score	AUC
SVM	0.806	0.00	0.00	0.00	0.50
RF	0.806	0.00	0.00	0.00	0.480
KNN	0.805	0.00	0.00	0.00	**0.510**
GB	**0.807**	**1.00**	**0.003**	**0.006**	0.475
XGBoost	0.806	0.00	0.00	0.00	0.490

As shown in Table 2, all models report an accuracy of approximately 80–81%, which at first glance may appear satisfactory. However, this accuracy is misleading and primarily reflects the models' tendency to predict only the majority class. This issue becomes evident when examining minority-class metrics. For Support Vector Machine, Random Forest, KNN, and XGBoost, the precision, recall, and F1-score for the chronic disease class are exactly zero, indicating a complete failure to identify any positive cases. The confusion matrices further confirm this behavior: nearly all chronic disease samples are misclassified as non-chronic, resulting in classifiers that effectively act as majority-class detectors. This outcome is a direct consequence of optimizing global loss functions under imbalanced conditions, where predicting the dominant class minimizes error without penalizing misclassifications of minority classes. Gradient Boosting exhibits a marginal deviation from this pattern by correctly identifying only one chronic disease instance, yielding a recall of 0.0034 and an F1-score of 0.0069. Despite its precision reaching 1.0 for the minority class, this value is statistically insignificant due to the extremely small number of true positive predictions. Moreover, the AUC values across all models remain close to 0.5, indicating performance equivalent to random guessing and reflecting the models' inability to meaningfully separate chronic and non-chronic classes under severe class imbalance.

Overall, the results in Table 2 demonstrate that modeling on imbalanced data leads to severe minority-class neglect, regardless of the underlying algorithmic complexity. Although ensemble and boosting-based models are theoretically robust, they remain fundamentally constrained by skewed class distributions. Consequently, without addressing class imbalance, the predictive models lack practical utility for chronic disease detection, underscoring the necessity of imbalance-aware techniques such as SMOTE–Tomek, which are addressed in the subsequent section.

B. Impact of SMOTE-Tomek

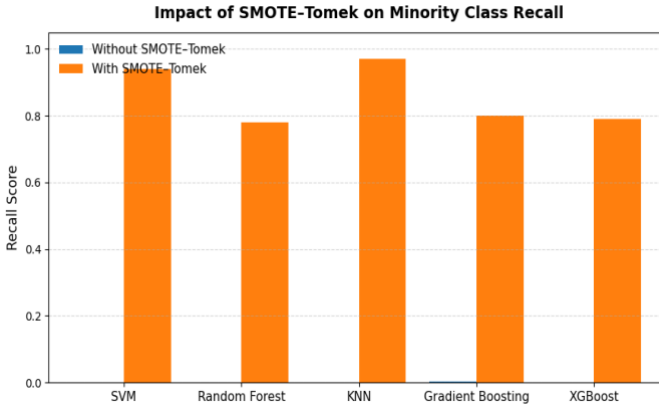

Fig. 3. Recall Comparison Across Models With and Without SMOTE–Tomek

To address the limitations observed in Section A, the dataset was reprocessed using the SMOTE–Tomek technique, which combines synthetic minority oversampling with Tomek Link undersampling to both increase minority representation and remove ambiguous majority samples. After this procedure, the dataset becomes perfectly balanced, with 4,789 samples per class, providing a more stable foundation for model learning. The effectiveness of SMOTE–Tomek is not only reflected in class balance but, more importantly, in substantial improvements in minority-class prediction performance. Unlike the imbalanced scenario, all

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

classifiers trained on the balanced dataset demonstrate dramatic gains in recall, F1-score, and AUC, indicating a restored ability to discriminate between chronic and non-chronic cases. To clearly illustrate the impact of SMOTE–Tomek, Figure 3 presents a bar-chart comparison of recall values for the chronic disease class across all models with and without SMOTE–Tomek. As shown in Figure 3, recall values increase from near-zero levels in the imbalanced setting to consistently high values after applying SMOTE–Tomek. Notably, KNN achieves the highest recall (0.9718), representing a dramatic increase from 0.00 in the imbalanced setting, while SVM improves from 0.00 to 0.9395. This substantial improvement demonstrates strong sensitivity to minority-class instances after applying SMOTE–Tomek. This improvement confirms that balancing the dataset fundamentally alters the models' learned decision boundaries, enabling them to capture minority-class patterns that were previously ignored.

Table 3. Performance of Classification Models With SMOTE–Tomek

Model	Acc	Preci	Recall	F1-Score	AUC
SVM	**0.929**	0.920	0.939	**0.929**	**0.981**
RF	0.847	0.906	0.775	0.835	0.902
KNN	0.804	0.728	**0.971**	0.832	0.898
GB	0.869	**0.924**	0.804	0.860	0.928
XGBoost	0.836	0.872	0.788	0.828	0.890

As shown in Table 3, among all evaluated models, SVM emerges as the best overall performer under the balanced setting, achieving the highest F1-score (0.9298) and AUC (0.9814), along with well-balanced precision (0.920) and recall (0.939). These metrics collectively indicate superior minority-class sensitivity and discriminative capability compared to other models. Importantly, these gains are not achieved at the expense of majority-class performance, as precision and recall remain well balanced across both classes. This demonstrates that SMOTE–Tomek does not merely inflate minority metrics but instead enables a more equitable and generalizable learning process.

IV. Discussion

When trained on the original imbalanced dataset, all evaluated models achieved accuracy values in the range of 0.805–0.807. However, minority-class performance was extremely poor, with recall equal to 0.00 for SVM, Random Forest, KNN, and XGBoost, while Gradient Boosting achieved only 0.0034 recall and an F1-score of 0.0069. The AUC values, ranging from 0.475 to 0.510, indicate near-random discrimination despite moderate overall accuracy. After applying SMOTE–Tomek, minority detection improved substantially across all models. SVM achieved 0.929 accuracy, 0.939 recall, 0.929 F1-score, and 0.981 AUC, while KNN reached the highest recall of 0.971. Random Forest, Gradient Boosting, and XGBoost achieved recall values of 0.775, 0.804, and 0.788, respectively, with corresponding AUC values above 0.89. These results indicate that imbalance handling significantly improved class separability and restored the models' ability to detect chronic disease cases.

After applying SMOTE–Tomek, minority detection improved substantially across all models. SVM achieved an accuracy of 0.929, a recall of 0.939, an F1-score of 0.929, and an AUC of 0.981. Compared to its pre-resampling recall of 0.00, this represents a 93.9% absolute increase in minority sensitivity. KNN achieved the highest recall (0.971), increasing from 0.00, although with lower precision (0.728) and overall accuracy (0.804). Random Forest improved to a recall of 0.775 (from 0.00) and an AUC of 0.902. Gradient Boosting increased recall from 0.0034 to 0.804, while XGBoost improved recall from 0.00 to 0.788, with an AUC of 0.890. These numerical changes demonstrate that SMOTE–Tomek does not merely adjust class proportions but significantly shifts decision boundaries to restore minority separability.

The superiority of SVM (AUC 0.981) under balanced conditions aligns partially with Kumar and Sikamani [19], who reported SVM accuracy of 91% on chronic disease datasets. In the present study, SVM accuracy increased to 92.9%, slightly higher than their reported value, suggesting that imbalance correction further enhances SVM robustness. However, compared with Ghosh et al. [18], who achieved 97% accuracy using Relief-based Random Forest, the present RF model achieved 84.7% accuracy and an AUC of 0.902. The lower RF accuracy may be attributed to differences in dataset composition, feature engineering strategy, and the use of synthetic data rather than curated cardiovascular datasets.

Saqlain et al. [20] reported SVM accuracies ranging from 81.19% to 92.68% across four UCI heart disease datasets. The present SVM accuracy (92.9%) lies within and slightly above their upper bound, while the AUC (0.981) indicates stronger class separability compared to their reported accuracy-focused evaluation. Mohan et al. [21] reported 88.7% accuracy using their hybrid HRFLM model. In contrast, the current SVM with imbalance handling achieved higher accuracy (92.9%) without requiring hybrid feature-model integration, indicating that class balancing alone can yield performance gains comparable to structural model modifications.

Yang et al. [22] reported XGBoost accuracy of 86.4% and AUC of 0.9139 using large-scale EHR data. In this study, XGBoost achieved 83.6% accuracy and an AUC of 0.890 after SMOTE–Tomek. While slightly lower than Yang et al.'s results, the performance remains comparable, despite using a synthetic dataset of 7,500 records rather than 42,000 real patient records. This suggests that class-imbalance handling contributes meaningfully to predictive discrimination even in non-clinical datasets.

From a methodological standpoint, the numerical evidence shows that recall improvement from 0.00 to values between 0.775 and 0.971 is critical for clinical relevance. In disease screening scenarios, missing

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

positive cases (false negatives) has higher cost than moderate reductions in precision. Therefore, imbalance-aware preprocessing should be considered mandatory in chronic disease prediction pipelines. Additionally, the finding that SVM achieved AUC 0.981 suggests that carefully balanced classical models can match or exceed more complex hybrid architectures reported in [21], provided that class distribution is properly managed.

From a practical perspective, the results indicate that healthcare decision-support systems should prioritize recall and AUC rather than accuracy alone. Without imbalance correction, models may achieve ~80% accuracy yet fail to detect any high-risk patients, rendering them unsuitable for deployment.

Despite its contributions, this study has several limitations that should be acknowledged. First, the dataset, while representative of lifestyle and health-related factors, is synthetically generated and may not fully capture real-world clinical noise, patient heterogeneity, or complex interactions such as genetic markers, imaging data, or longitudinal disease progression, which may affect direct generalizability to clinical environments. Second, SMOTE–Tomek generates synthetic samples based on existing data distributions, which may introduce bias if the original minority-class patterns are not sufficiently diverse. Future work may mitigate this limitation by incorporating real clinical datasets, applying more advanced generative resampling techniques, or validating synthetic samples through domain expert review. Third, the analysis focuses on traditional machine learning classifiers; future work could explore the interaction between methods for handling class imbalance and deep learning architectures.

V. Conclusion

This study aimed to evaluate the impact of SMOTE–Tomek resampling on the classification performance of chronic disease prediction models using health and lifestyle features. The experimental results demonstrate that models trained on the original imbalanced dataset achieved relatively high accuracy (approximately 80–85%) but exhibited extremely low recall for the minority class, indicating poor detection capability for chronic disease cases. After applying SMOTE–Tomek, all models showed substantial improvements in minority-class detection performance. In particular, Support Vector Machine (SVM) achieved the best overall performance, reaching an accuracy above 90%, with significant increases in recall, F1-score, and AUC compared to the imbalanced scenario. Random Forest and XGBoost also demonstrated competitive improvements, while KNN and LSTM benefited from resampling but did not outperform SVM. These findings confirm that imbalance handling is a decisive factor in chronic disease classification and that evaluation metrics such as recall and F1-score must be prioritized over accuracy in imbalanced medical datasets.

For future work, validation using real-world clinical datasets is necessary to assess generalizability beyond synthetic data. Further investigation may also explore advanced resampling techniques or generative models to enhance minority-class representation, as well as the integration of longitudinal health records and multimodal clinical features. Additionally, combining imbalance handling with hybrid feature selection and advanced hyperparameter optimization strategies may further improve predictive robustness and clinical applicability.

Overall, this research advances the field by empirically validating that effective treatment of class imbalance is a fundamental requirement for building clinically meaningful machine learning models, rather than a secondary optimization step. The proposed framework can be applied to chronic disease screening systems that rely on routinely collected health and lifestyle data. Future work should focus on validating this approach on larger real-world clinical datasets, incorporating longitudinal information, and exploring hybrid predictive architectures (e.g., combinations of machine learning and deep learning models), to further enhance predictive robustness and generalizability.

References

[1] Diet, W., 2003. Chronic diseases. Geneva: World Health Organization

[2] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, “A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease,” Decision Analytics Journal, vol. 6, p. 100169, Mar. 2023, doi: https://doi.org/10.1016/j.dajour.2023.100169.

[3] H. A. Al-Jamimi, “Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction,” IEEE Access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3395512.

[4] R. Islam, A. Sultana, and Mohammad Rashedul Islam, “A comprehensive review for chronic disease prediction using machine learning algorithms,” Journal of Electrical Systems and Information Technology, vol. 11, no. 1, Jul. 2024, doi: https://doi.org/10.1186/s43067-024-00150-4.

[5] Y. Yan and J. Mi, “Noncommunicable chronic disease prevention should start from childhood,” Pediatric Investigation, vol. 5, no. 1, pp. 3–5, Mar. 2021, doi: https://doi.org/10.1002/ped4.12254.

[6] World Health Organization (2005) WHO steps surveillance manual : the WHO stepwise approach to chronic disease risk factor surveillance.

[7] S. Nusinovici et al., “Logistic regression was as good as machine learning for predicting major chronic diseases,” Journal of Clinical Epidemiology, vol. 122, pp. 56–69, Jun. 2020, doi: https://doi.org/10.1016/j.jclinepi.2020.03.002.

[8] 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, doi: https://doi.org/10.1109/access.2021.3080316.

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

[9] Kamaladevi M, Venkataraman, and S. K. R, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," Annals of the Romanian Society for Cell Biology, Apr. 2021.

[10] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," Journal of Biomedical Informatics, vol. 107, p. 103465, Jul. 2020, doi: https://doi.org/10.1016/j.jbi.2020.103465.

[11] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," JOIV : International Journal on Informatics Visualization, vol. 7, no. 1, pp. 258–264, Feb. 2023, doi: https://doi.org/10.30630/joiv.7.1.1069.

[12] None Shahmirul Hafizullah Imanuddin, None Kusworo Adi, and None Rahmat Gernowo, "Sentiment Analysis on Satusehat Application Using Support Vector Machine Method," Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 5, no. 3, pp. 143–149, Jul. 2023, doi: https://doi.org/10.35882/jeemi.v5i3.304.

[13] Dhiyaussalam, A. Wibowo, F. A. Nugroho, E. A. Sarwoko, and I. M. A. Setiawan, "Classification of Headache Disorder Using Random Forest Algorithm," IEEE Xplore, Nov. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9299105

[14] G. Mitrentsis and H. Lens, "An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting," Applied Energy, vol. 309, p. 118473, Mar. 2022, doi: https://doi.org/10.1016/j.apenergy.2021.118473.

[15] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of Gradient Boosting Algorithms," Artificial Intelligence Review, vol. 54, no. 3, Aug. 2020, doi: https://doi.org/10.1007/s10462-020-09896-5.

[16] Y. Tang, Y.-C. Chang, and K. Li, "Applications of K-nearest neighbor algorithm in intelligent diagnosis of wind turbine blades damage," Renewable Energy, vol. 212, pp. 855–864, Aug. 2023, doi: https://doi.org/10.1016/j.renene.2023.05.087.

[17] Q. Lin, B. Lin, D. Zhang, and J. Wu, "Web-based prototype system for flood simulation and forecasting based on the HEC-HMS model," Environmental Modelling and Software, vol. 158, pp. 105541–105541, Dec. 2022, doi: https://doi.org/10.1016/j.envsoft.2022.105541.

[18] P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," IEEE Access, vol. 9, pp. 19304–19326, 2021, doi: https://doi.org/10.1109/access.2021.3053759.

[19] N. Kumar and K. Sikamani, "Prediction of Chronic and Infectious Diseases using Machine Learning Classifiers- A Systematic Approach," International Journal of Intelligent Engineering and Systems, vol. 13, no. 4, pp. 11–20, Aug. 2020, doi: https://doi.org/10.22266/ijies2020.0831.02.

[20] S. M. Saqlain et al., "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowledge and Information Systems, vol. 58, no. 1, pp. 139–167, Mar. 2018, doi: https://doi.org/10.1007/s10115-018-1185-y.

[21] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, no. 1, pp. 81542–81554, 2019, doi: https://doi.org/10.1109/access.2019.2923707.

[22] Y. Yang, Y. Li, R. Chen, J. Zheng, Y. Cai, and G. Fortino, "Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data," Big Data Research, vol. 25, p. 100234, Jul. 2021, doi: https://doi.org/10.1016/j.bdr.2021.100234.

[23] A. Mousa, F. Özyurt, and E. Avcı, "Enhancing Credit Card Fraud Detection Using Synthetic Minority Over-Sampling Technique (SMOTE) and Deep Neural Networks: A Comprehensive Analysis," International Journal of Advanced Networking and Application, vol. 16, no. 03, pp. 6390–6401, 2024, doi: https://doi.org/10.35444/ijana.2024.16304.

[24] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," Sensors, vol. 22, no. 9, p. 3246, Jan. 2022, doi: https://doi.org/10.3390/s22093246.

[25] A. Khleel and Károly Nehéz, "A novel approach for software defect prediction using CNN and GRU based on SMOTE Tomek method," Journal of Intelligent Information Systems, vol. 60, no. 3, pp. 673–707, May 2023, doi: https://doi.org/10.1007/s10844-023-00793-1.

[26] V. R. Joseph, "Optimal ratio for data splitting," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 15, no. 4, pp. 531–538, Apr. 2022, doi: https://doi.org/10.1002/sam.11583.

[27] Tajali, A., Saragih, T.H., Mazdadi, M.I., Budiman, I. and Farmadi, A., 2024. The Impactness of SMOTE as Imbalance Class Handling for Myocardial Infarction Complication Classification using Machine Learning Approach with Data Imputation and Hyperparameter. Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, 6(4), pp.227-239.

[28] S. Talukdar et al., "Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review," Remote Sensing, vol. 12, no. 7, p. 1135, Apr. 2020, doi: https://doi.org/10.3390/rs12071135.

[29] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal: Promoting Communications on Statistics

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

and Stata, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: https://doi.org/10.1177/1536867x20909688.
[30] D. C. E. Saputra, Y. Maulana, T. A. Win, R. Phann, and W. Caesarendra, "Implementation of Machine Learning and Deep Learning Models Based on Structural MRI for Identification Autism Spectrum Disorder," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, vol. 9, no. 2, pp. 307–318, May 2023, doi: https://doi.org/10.26555/jiteki.v9i2.26094.
[31] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interactive Learning Environments, vol. 31, no. 6, pp. 1–20, May 2021, doi: https://doi.org/10.1080/10494820.2021.1928235.
[32] T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: https://doi.org/10.1145/2939672.2939785.
[33] Y. Hamed, A. Ibrahim Alzahrani, A. Shafie, Z. Mustaffa, M. Che Ismail, and K. Kok Eng, "Two steps hybrid calibration algorithm of support vector regression and K-nearest neighbors," Alexandria Engineering Journal, vol. 59, no. 3, pp. 1181–1190, Jun. 2020, doi: https://doi.org/10.1016/j.aej.2020.01.033.
[34] B. Du, S. Wang, N. Wang, L. Zhang, D. Tao, and L. Zhang, "Hyperspectral signal unmixing based on constrained non-negative matrix factorization approach," Neurocomputing (Amsterdam), vol. 204, pp. 153–161, Sep. 2016, doi: https://doi.org/10.1016/j.neucom.2015.10.132.
[35] H. A. Abu Alfeilat et al., "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review," Big Data, vol. 7, no. 4, pp. 221–248, Dec. 2019, doi: https://doi.org/10.1089/big.2018.0175.
[36] I. Lee and C. Torpelund-Bruin, "Geographic knowledge discovery from Web Map segmentation through generalized Voronoi diagrams," Expert Systems with Applications, vol. 39, no. 10, pp. 9376–9388, Aug. 2012, doi: https://doi.org/10.1016/j.eswa.2012.02.129.
[37] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," ACM Computing Surveys, vol. 54, no. 6, pp. 1–25, Jul. 2021, doi: https://doi.org/10.1145/3459665.
[38] Y. Tang, Y.-C. Chang, and K. Li, "Applications of K-nearest neighbor algorithm in intelligent diagnosis of wind turbine blades damage," Renewable Energy, vol. 212, pp. 855–864, Aug. 2023, doi: https://doi.org/10.1016/j.renene.2023.05.087.
[39] Jung Min Ahn, J. Kim, and K. Kim, "Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting," Toxins, vol. 15, no. 10, pp. 608–608, Oct. 2023, doi: https://doi.org/10.3390/toxins15100608.
[40] T. Sathish, Divity SaiKumar, S. Patil, R. Saravanan, J. Giri, and A. A. Aly, "Exponential smoothing method against the gradient boosting machine learning algorithm-based model for materials forecasting to minimize inventory," AIP Advances, vol. 14, no. 6, Jun. 2024, doi: https://doi.org/10.1063/5.0208491.

Author Biography

Muhammad Adika Riswanda received his undergraduate education in Computer Science at Lambung Mangkurat University, Indonesia, where he has been enrolled since 2020. His academic interests primarily focus on data science, machine learning, and data-driven problem-solving, particularly in the application of computational methods to health and predictive analytics research. His research activities include data preprocessing, handling class imbalance, and developing and evaluating machine learning models for classification tasks. He can be contacted at email: 2011016210025@mhs.ulm.ac.id

Friska Abadi finished his bachelor's degree in Computer Science from Lambung Mangkurat University in 2011. Subsequently, in 2016, he obtained his master's degree from the Department of Informatics at STMIK Amikom, Yogyakarta. Following that, he joined Lambung Mangkurat University as a lecturer in Computer Science. As a lecturer, he teaches programming. Apart from that, he also carries out research and community service. Other activities as an application developer, whether using a web or mobile platform. Currently, he holds the position of head of the software engineering laboratory. His current area of research revolves around software engineering, and he is also interested in machine learning. He can be contacted at email: friska.abadi@ulm.ac.id.

Muhammad Itqan Mazdadi, a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. Before becoming a lecturer, he completed his undergraduate program in the Computer Science Departmentat Lambung Mangkurat University in 2013. He then completed his master's degree from the Department of Informatics at Islamic University of Indonesia, Yogyakarta. Currently, he serves as the Secretary of the Computer Science Department at Lambung Mangkurat University. He can be contacted at email: mazdadi@ulm.ac.id.

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324

Mohammad Reza Faisal was born in Banjarmasin. After graduating from high school, he pursued his undergraduate studies in the Department of Informatics at Universitas Pasundan in 1995 and later majored in Physics at Institut Teknologi Bandung (ITB) in 1997. After completing his bachelor's degree, he gained professional experience as an information technology and software development trainer. Since 2008, he has been a lecturer in Computer Science at Universitas Lambung Mangkurat. He completed his master's degree in Informatics at Institut Teknologi Bandung in 2010. In 2015, he earned a doctoral degree in Bioinformatics from Kanazawa University, Japan. He continues to serve as a lecturer in Computer Science at Universitas Lambung Mangkurat. His research interests include Data Science, Software Engineering, and Bioinformatics.

Email: reza.faisal@ulm.ac.id

Rudy Herteno received his bachelor's degree in Computer Science from Lambung Mangkurat Universityin 2011. After completing his studies, he worked as a software developer for several years to gain more experience in the field. During this period, he developed various software applications, particularly to support the needs of local governments. In 2017, he obtained a master's degree in Informatics from STMIK Amikom University. Currently, he is a lecturer in the Computer Science program at Lambung Mangkurat University. His research interests include software engineering, software defect prediction, and deep learning, aiming to improve software quality, optimize error detection in systems, and develop artificial intelligence-based solutions. He can be contacted at email: rudy.herteno@ulm.ac.id

Corresponding author: Friska Abadi, friska.abadi@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i1.324