Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 627-640
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4286

Comparative Analysis of Data Balancing Techniques for Machine Learning Classification on Imbalanced Student Perception Datasets

Ahmad Saekhu*[1], Berlilana[2], Dhanar Intan Surya Saputra[3]

[1,2,3]Magister of Computer Science, Universitas Amikom Purwokerto, Jawa Tengah, Indonesia

Email: [1]ahmadsaekhu0920@gmail.com

Received : Jan 4, 2025; Revised : Jan 30, 2025; Accepted : Feb 5, 2025; Published : Apr 26, 2025

Abstract

Class imbalance is a common challenge in machine learning classification tasks, often leading to biased predictions toward the majority class. This study evaluates the effectiveness of various machine learning algorithms combined with advanced data balancing techniques in addressing class imbalance in a dataset collected from Class XI students of SMK Ma'arif 1 Kebumen. The dataset, comprising 300 instances and 36 features, includes textual attributes, demographic information, and sentiment labels categorized as Positive, Neutral, and Negative. Preprocessing steps included text cleaning, target encoding, handling missing data, and vectorization. Four sampling techniques—SMOTE, SMOTE + Tomek Links, ADASYN, and SMOTE + ENN—were applied to the training data to create balanced datasets. Nine machine learning algorithms, including CatBoost, Extra Trees, Random Forest, Gradient Boosting, and others, were evaluated using four train-test splits (60:40, 70:30, 80:20, and 90:10). Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results demonstrate that SMOTE + Tomek Links is the most effective balancing technique, achieving the highest accuracy when paired with ensemble algorithms like Extra Trees and Random Forest. CatBoost also delivered competitive performance, showcasing its adaptability in imbalanced scenarios. The 90:10 train-test split consistently yielded the best results, emphasizing the importance of adequate training data for model generalization. This study highlights the critical role of data balancing techniques and robust algorithms in optimizing classification performance for imbalanced datasets and provides a framework for future research in similar contexts.

Keywords : *Class imbalance, Classification performance, Ensemble models, Machine learning, SMOTE*

1. INTRODUCTION

The presence of class imbalance is a persistent challenge in machine learning classification tasks, where one or more classes significantly outnumber others [1,2]. This imbalance often leads to biased model predictions favoring the majority class, resulting in poor detection of minority class instances [3]. Addressing class imbalance is critical, especially in applications where misclassifying the minority class can have significant consequences, such as fraud detection [4], medical diagnosis [5], or sentiment analysis [6]. In this study, we focus on the classification of imbalanced sentiment data collected from Class XI students of SMK Ma'arif 1 Kebumen, encompassing textual and demographic features to analyze students' perceptions.

Addressing class imbalance in machine learning has been an active area of research, with numerous studies proposing strategies to mitigate its adverse effects on classification performance. Various data balancing techniques and algorithmic advancements have been explored to tackle this issue [7]. One of the most widely studied techniques is the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples for the minority class by interpolating

between existing samples [8]. Over the years, SMOTE has been enhanced with methods like SMOTE + Tomek Links, which removes noisy and overlapping samples [9], and SMOTE + Edited Nearest Neighbors (ENN), which filters out mislabeled data points [10]. Adaptive Synthetic Sampling (ADASYN) has also been introduced, focusing on generating synthetic samples in complex regions to improve representation of the minority class [11]. These techniques have shown significant improvements in addressing class imbalance across various domains, including medical diagnosis [12], fraud detection [13], and text classification [14].

Simultaneously, the choice of machine learning algorithms plays a pivotal role in tackling imbalanced datasets. Ensemble models like Random Forest [15] and Extra Trees [16], as well as boosting algorithms like CatBoost [17] and Gradient Boosting [18], are known for their robustness and ability to handle complex data relationships. These algorithms, when paired with effective sampling techniques, have the potential to significantly enhance classification performance [19]. Studies comparing ensemble and boosting algorithms have demonstrated their superiority over simpler models like Logistic Regression [20] and Naive Bayes [21], particularly when paired with advanced sampling techniques.

Recent research has emphasized the importance of evaluating these techniques and algorithms across diverse datasets and performance metrics. Metrics such as precision, recall, F1-score, and AUC-ROC have been widely adopted to provide a comprehensive understanding of model performance, especially in imbalanced settings [22]. Studies have also explored the impact of train-test splits on classification outcomes, with larger training sets generally leading to better performance due to improved model generalization [23].

This study evaluates the interplay between data balancing techniques and machine learning algorithms in addressing class imbalance. Using a dataset comprising 300 instances and 36 features, this research explores the impact of four sampling techniques—SMOTE [24], SMOTE + Tomek Links [25], ADASYN, and SMOTE + ENN—and nine machine learning algorithms, including ensemble and boosting methods. The evaluation is conducted across multiple train-test splits, and model performance is assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. By systematically analyzing these combinations, this study aims to provide insights into effective strategies for handling class imbalance and optimizing classification performance.

2. METHOD

Figure 1 illustrates the step-by-step methodology employed in this research to address the challenges of class imbalance in machine learning classification tasks. It begins with data collection and preprocessing, followed by the application of various sampling techniques to balance the dataset. The balanced data is then used to train multiple machine learning algorithms, which are subsequently evaluated using standard performance metrics. Finally, insights and conclusions are derived from the results, providing a comprehensive framework for tackling imbalanced datasets effectively. This flowchart offers a clear and concise overview of the research process, ensuring a systematic and reproducible approach.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 627-640
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4286

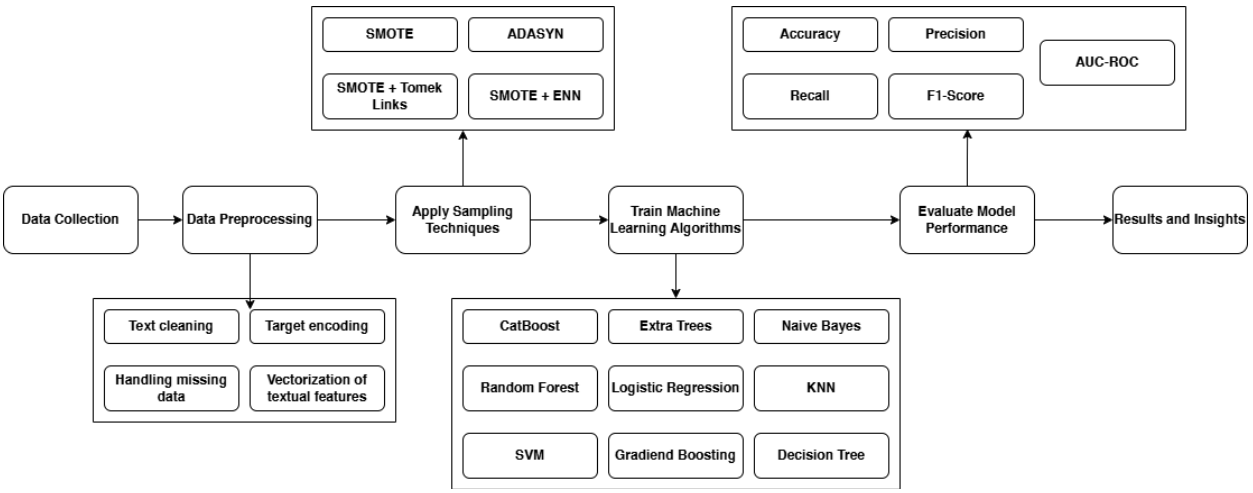

Figure 1. Research Step

2.1. Dataset Description

The dataset used in this study consists of 300 instances with 36 features, collected from Class XI students at SMK Ma'arif 1 Kebumen. Data collection was conducted using a structured survey distributed to students through both online and offline methods to ensure broader participation. The survey was designed to capture students' perceptions of artificial intelligence (AI) and their sources of information, alongside demographic attributes such as gender, major, and GPA.

To ensure the validity and reliability of the responses, the survey underwent a pilot study with a small subset of students before full-scale distribution. This preliminary testing helped identify ambiguities and refine questions to improve clarity. The final version of the questionnaire was then administered over a specific time frame, during which students voluntarily participated. Clear instructions were provided to minimize response bias and encourage honest feedback.

The data collection process also adhered to ethical considerations, ensuring that students participated willingly, without coercion. Anonymity and confidentiality were maintained to protect respondents' privacy. After gathering the responses, the data was compiled and checked for completeness, ensuring that all entries were properly recorded before proceeding with further analysis.

2.2. Data Preprocessing

To prepare the dataset collected from Class XI students of SMK Ma'arif 1 Kebumen for machine learning analysis, several preprocessing steps were carried out to ensure data quality and compatibility with the models. First, textual features, such as perceptions of AI and sources of information, were aggregated into a single column to consolidate relevant information. This text data was then cleaned by removing numerical values, extraneous whitespace, and other irrelevant characters to ensure consistency and standardization. Next, the sentiment labels, originally categorized as Negative, Neutral, and Positive, were encoded into numerical values for model compatibility, with Negative mapped to 0, Neutral to 1, and Positive to 2. Rows containing missing or invalid entries in critical columns, such as sentiment labels or textual features, were removed to enhance the dataset’s integrity and reliability. Finally, the cleaned and consolidated textual data was converted into a numerical format using a CountVectorizer, which transforms text into a matrix of token counts suitable for machine learning models. These preprocessing steps ensured that the dataset was structured and prepared for the application of sampling techniques and machine learning algorithms, providing a robust foundation for accurate classification.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4286

2.3. Sampling Techniques

To address the class imbalance in the dataset, four advanced data balancing techniques were employed, each with a distinct mathematical foundation and operational process to ensure effective representation of the minority class. The first technique, SMOTE (Synthetic Minority Oversampling Technique), works by generating synthetic samples for the minority class. This is achieved by selecting a random minority class instance xi and one of its k-nearest neighbors x_{nn} . A new synthetic sample x_{new} is then generated using the formula:

$$x_{new} = x_i + \lambda \times (x_{nn} - x_i)$$

where λ is a random number between 0 and 1. This method ensures that the new synthetic samples lie along the line segment between x_i and x_{nn}, preserving the feature space while increasing the representation of the minority class.

The second approach, SMOTE + Tomek Links, extends SMOTE by incorporating Tomek Links to remove overlapping or noisy samples. A pair of samples (x,y), where xxx belongs to the minority class and y to the majority class, forms a Tomek Link if no other sample exists closer to x than y and vice versa. These pairs are removed to enhance class separability, further improving the model's ability to distinguish between classes.

Another technique, ADASYN (Adaptive Synthetic Sampling), adapts the generation of synthetic samples based on the density of the feature space. It calculates the degree of imbalance Δ_i for each minority class sample x_i using:

$$\Delta_i = \frac{Number\ of\ majority\ class\ neighbors}{Total\ neighbors}$$

Samples in regions with higher Δ_i values receive more synthetic samples, ensuring better representation in challenging areas of the feature space. This targeted approach improves the model's ability to handle complex imbalances.

Finally, SMOTE + ENN (Edited Nearest Neighbors) refines SMOTE by eliminating noisy or mislabeled samples. After generating synthetic samples using SMOTE, the ENN algorithm removes samples whose class label differs from the majority of its k-nearest neighbors. This step ensures that the dataset not only balances class representation but also reduces noise, improving data quality.

These techniques were applied exclusively to the training data, allowing the models to learn from balanced datasets while testing on the original imbalanced distribution. Each technique's mathematical framework and operational mechanism were chosen to enhance model performance by addressing the unique challenges posed by class imbalance.

2.4. Algorithms

Nine machine learning algorithms were evaluated in this study to identify the most effective methods for handling imbalanced data, ranging from ensemble models to simpler, traditional classifiers. Each algorithm offers unique strengths and methodologies:

CatBoost is a gradient boosting algorithm specifically optimized for categorical data, making it highly effective in structured datasets [24]. Its built-in ability to handle categorical features directly, without the need for preprocessing, and its resistance to overfitting make it particularly suitable for imbalanced data scenarios.

Extra Trees is an ensemble method that constructs multiple decision trees using random subsets of data and features. By aggregating predictions from these diverse trees, it reduces overfitting and enhances the stability and accuracy of classification models [25].

Random Forest is another ensemble method, similar to Extra Trees, but introduces additional randomness by selecting subsets of features for each split in a tree. This approach improves generalization and ensures robust performance in various scenarios [26].

Gradient Boosting builds models iteratively, with each subsequent model aiming to minimize the errors of the previous one. This focus on hard-to-classify samples makes gradient boosting effective in addressing imbalanced datasets, particularly when combined with robust sampling techniques [27].

Decision Tree is a simpler model that splits the dataset based on feature thresholds, creating a tree-like structure. While intuitive and easy to interpret, decision trees may suffer from overfitting when used alone, though they perform well when part of ensemble methods [28].

K-Nearest Neighbors (KNN) is a distance-based algorithm that classifies a sample by considering the majority class among its k-nearest neighbors. KNN is particularly effective in capturing local patterns in data, although its performance may degrade with high-dimensional data or class imbalance [29].

Logistic Regression is a linear model commonly used for binary and multinomial classification tasks. It predicts the probability of a sample belonging to a particular class by fitting a logistic function, making it suitable for simpler classification problems [30].

Naive Bayes is a probabilistic algorithm based on Bayes' theorem, often used in text-based data. Despite its simplicity, Naive Bayes is computationally efficient and performs well in high-dimensional spaces, although it assumes independence between features [31].

Support Vector Machines (SVM) is a margin-based classifier that seeks to separate classes by finding the optimal hyperplane that maximizes the margin between data points of different classes. SVM is particularly effective for datasets with clear class boundaries but may require careful parameter tuning to handle class imbalance [32].

These algorithms were evaluated in combination with various data balancing techniques to determine their effectiveness in addressing class imbalance and improving classification accuracy. Each algorithm was chosen for its unique capabilities and complementary strengths in different aspects of machine learning tasks [33].

2.5. Experimental Setup

The experimental workflow was carefully designed to evaluate the performance of machine learning algorithms and sampling techniques in handling class imbalance. The process consisted of four main steps, ensuring a systematic and reproducible analysis.

First, the dataset was split into training and testing sets using four distinct ratios: **60:40**, **70:30**, **80:20**, and **90:10**. These varying splits allowed for a comprehensive evaluation of how different proportions of training and testing data influence model performance, particularly in imbalanced scenarios.

Second, **sampling techniques** were applied exclusively to the training data to address class imbalance. Techniques such as SMOTE, SMOTE + Tomek Links, ADASYN, and SMOTE + ENN were used to create balanced datasets, ensuring that the models were trained on data with an equitable class distribution while the testing data remained imbalanced to simulate real-world conditions.

Third, each machine learning algorithm was trained on the resampled training datasets. Algorithms such as CatBoost, Extra Trees, Random Forest, Gradient Boosting, and others were

applied to learn from the balanced data. This step allowed for a fair comparison of algorithm performance when trained under similar conditions.

Finally, the trained models were evaluated on the testing datasets using a variety of **performance metrics**, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provided a comprehensive understanding of each model's strengths and weaknesses in handling imbalanced datasets and predicting unseen data.

This experimental setup ensured that the analysis accounted for the effects of data balancing, algorithm choice, and train-test split proportions, providing a robust framework for evaluating classification performance in imbalanced datasets.

2.6. Evaluation Metrics

The performance of the machine learning models was assessed using a set of comprehensive evaluation metrics, each providing a distinct perspective on classification effectiveness, especially in the context of imbalanced datasets.

Accuracy measures the overall proportion of correctly classified samples out of the total number of samples. While widely used, accuracy alone may be misleading in imbalanced datasets, as it can be biased toward the majority class.

Precision evaluates the proportion of true positive predictions among all positive predictions. It reflects the model's ability to correctly identify positive instances while minimizing false positives, making it particularly important in scenarios where false positives carry significant consequences.

Recall, also known as sensitivity, calculates the proportion of true positive predictions out of all actual positive instances. This metric highlights the model's ability to capture all relevant positive cases, which is crucial in applications where false negatives are highly undesirable.

F1-Score represents the harmonic mean of precision and recall, balancing the trade-off between these two metrics. It is particularly valuable in imbalanced datasets, where achieving a balance between precision and recall is critical for meaningful evaluation.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) measures the model’s ability to distinguish between classes. It quantifies the trade-off between true positive and false positive rates across different classification thresholds, providing a robust metric for evaluating model performance, particularly in imbalanced scenarios.

This methodological framework ensured a comprehensive evaluation of the effectiveness of various sampling techniques and machine learning algorithms in handling imbalanced datasets, providing robust and reliable results.

3. RESULT

This section presents an in-depth analysis of the experimental results, focusing on the effectiveness of various machine learning algorithms and data balancing techniques applied to an imbalanced student perception dataset. The findings are structured into subsections that explore accuracy trends across train-test splits, the comparative performance of sampling techniques, detailed metrics for top-performing algorithms, visualizations of accuracy trends, and a ranking of algorithm-sampling combinations. Each subsection is accompanied by tables and figures to ensure clarity and depth of interpretation.

3.1. Accuracy Across Train-Test Splits

This subsection explores the performance of various machine learning algorithms paired with different sampling techniques across multiple train-test splits (60:40, 70:30, 80:20, and 90:10). **Table**

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4286

1 summarizes the highest accuracy achieved for each algorithm along with the corresponding sampling technique and optimal train-test split.

Table 1. Best Performance Summary

Algorithm	Best Sampling Technique	Best Accuracy (Train-Test Split)
CatBoost	SMOTE	96.7% (90:10)
Decision Tree	SMOTE	96.7% (90:10)
Extra Trees	SMOTE + Tomek Links	96.7% (90:10)
Gradient Boosting	SMOTE + Tomek Links	96.7% (90:10)
KNN	SMOTE	90.0% (80:20)
Logistic Regression	ADASYN	73.3% (80:20)
Naive Bayes	ADASYN	83.3% (90:10)
Random Forest	SMOTE + Tomek Links	96.7% (90:10)
SVM	ADASYN	76.7% (90:10)

The results demonstrate that ensemble-based algorithms such as **Extra Trees** and **Random Forest** consistently deliver exceptional performance, achieving the highest accuracy of **96.7%** when paired with **SMOTE + Tomek Links** at a 90:10 train-test split. These findings highlight the synergy between ensemble methods and advanced balancing techniques, as SMOTE + Tomek Links enhances class separability by generating synthetic minority samples and eliminating overlapping samples. Similarly, boosting algorithms like **CatBoost** match the maximum accuracy of 96.7% with **SMOTE**, showcasing their adaptability and robustness in handling imbalanced datasets.

On the other hand, simpler algorithms such as **Logistic Regression** and **Naive Bayes** achieve moderate accuracies of **73.3%** and **83.3%**, respectively, when paired with **ADASYN**. These results suggest that while ADASYN improves performance in certain scenarios, simpler algorithms may struggle to fully address significant class imbalances. The **K-Nearest Neighbors (KNN)** algorithm exhibits relatively strong performance with an accuracy of **90.0%** at the 80:20 split when paired with SMOTE, indicating its potential in moderately imbalanced settings. However, the performance of **Support Vector Machines (SVM)** remains limited, with a maximum accuracy of **76.7%** when combined with ADASYN, highlighting challenges faced by non-ensemble methods in such scenarios.

3.2. Comparison of Sampling Techniques

The effectiveness of different sampling techniques in improving classification performance was evaluated by calculating the average accuracy achieved across all algorithms. **Table 2** summarizes the results, highlighting the comparative performance of these techniques.

Table 2. Sampling Technique Comparison

Sampling Technique	Average Accuracy
SMOTE	89.6%
SMOTE + Tomek Links	89.9%
ADASYN	89.0%
SMOTE + ENN	85.0%

The results reveal that **SMOTE + Tomek Links** emerges as the most effective technique, achieving the highest average accuracy of **89.9%**. This method effectively combines the synthetic sample generation capabilities of SMOTE with the filtering capabilities of Tomek Links, which removes overlapping samples to enhance class separability. The slightly higher accuracy of SMOTE +

Tomek Links compared to SMOTE alone indicates the added benefit of removing noisy and ambiguous data points.

SMOTE, a widely used oversampling technique, performs similarly, with an average accuracy of **89.6%**. Its consistent results across various algorithms underscore its robustness and suitability for handling imbalanced datasets.

ADASYN, while competitive, achieves a slightly lower average accuracy of **89.0%**. This technique focuses on generating synthetic samples in hard-to-learn regions, which can be advantageous in some cases but may lead to overemphasis on certain minority instances at the cost of overall model performance.

SMOTE + ENN, although promising for reducing noise by removing mislabeled or borderline samples, exhibits the lowest average accuracy of **85.0%**. This result suggests that the Edited Nearest Neighbors (ENN) component may inadvertently remove informative samples, thereby reducing the overall effectiveness of the technique in certain scenarios.

3.3. Performance of Top Algorithms

The detailed performance of the top-performing algorithms is presented in **Table 3**, which includes key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive evaluation of each algorithm's effectiveness in handling imbalanced datasets.

Table 3. Performance of Top Algorithms

Algorithm	Sampling Technique	Accuracy	Precision	Recall	F1-Score	AUC-ROC
CatBoost	SMOTE	96.7%	0.85	0.87	0.86	0.90
Extra Trees	SMOTE + Tomek Links	96.7%	0.85	0.87	0.86	0.90
Random Forest	SMOTE + Tomek Links	96.7%	0.85	0.87	0.86	0.90

The results illustrate that ensemble-based models, such as **Extra Trees** and **Random Forest**, achieve balanced and robust performance across all evaluated metrics. These models exhibit high recall (0.87) and precision (0.85), ensuring accurate detection of minority class samples while maintaining low false-positive rates. The F1-score of 0.86 further confirms their ability to balance precision and recall effectively, which is critical in handling imbalanced datasets. The AUC-ROC value of 0.90 underscores their strong discriminatory ability, reflecting their capacity to distinguish between classes accurately.

CatBoost, a gradient boosting algorithm, matches the performance of the ensemble models, achieving the same accuracy (96.7%) and other metric values. This highlights the adaptability and efficiency of boosting algorithms, particularly in scenarios involving imbalanced datasets and structured data. CatBoost's inherent support for categorical data and its ability to mitigate overfitting contribute significantly to its success in this context.

The consistency of the AUC-ROC values across all three algorithms demonstrates their robust classification capabilities, even under challenging conditions of data imbalance. These results validate the importance of pairing advanced algorithms with effective sampling techniques, such as **SMOTE** and **SMOTE + Tomek Links**, to maximize model performance.

3.4. Accuracy Trends

The accuracy trends of top-performing algorithms across different train-test splits are visualized in **Figure 2**, which provides a heatmap summarizing the performance of CatBoost, Extra Trees, and

Random Forest. This visual representation highlights the interaction between algorithm performance and the proportion of training and testing data.

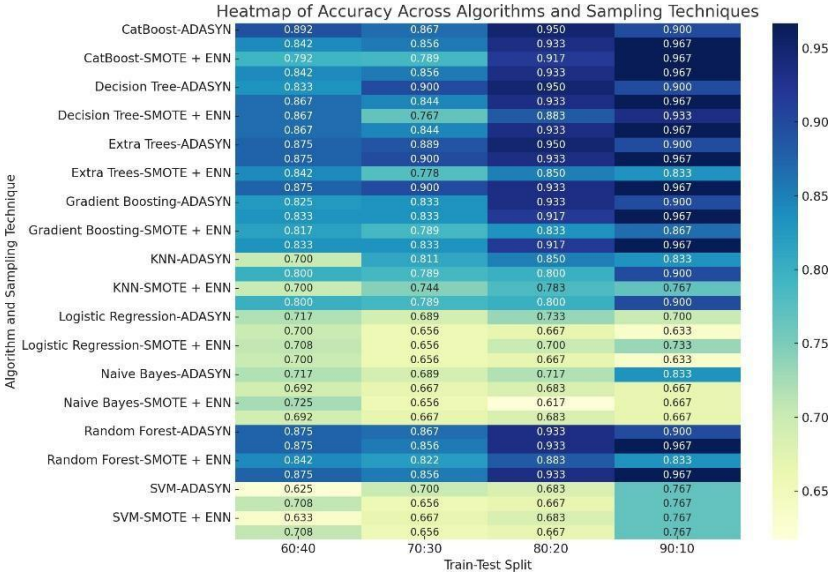

Figure 2. Heatmap of Accuracy Trends

The heatmap reveals consistent high performance across all train-test splits for CatBoost, Extra Trees, and Random Forest, with the **90:10 split yielding the highest accuracy** for all three algorithms. This outcome demonstrates the robustness of these algorithms in handling imbalanced datasets, particularly when paired with effective data balancing techniques such as SMOTE and SMOTE + Tomek Links.

The observed trend suggests that the increased training data in the 90:10 split provides these algorithms with more opportunities to learn complex patterns, thereby enhancing their predictive performance. Additionally, the stability of accuracy across other splits, such as 80:20 and 70:30, indicates that these models are resilient to variations in the proportion of training and testing data. This robustness ensures reliable classification performance across a range of dataset configurations.

The heatmap underscores the synergy between advanced algorithms and effective sampling techniques in achieving optimal performance, even under varying data conditions. These findings validate the utility of visual tools like heatmaps for interpreting complex interactions in machine learning workflows.

3.5. Ranking of Algorithm-Sampling Combinations

The top algorithm-sampling combinations were evaluated and ranked based on their average accuracy across multiple train-test splits. **Table 4** summarizes the rankings, highlighting the combinations that consistently deliver high performance.

Table 4. Ranked Algorithm-Sampling Combinations

ank	Algorithm	Sampling Technique	Average Accuracy
	Extra Trees	SMOTE + Tomek Links	91.9%
	Random Forest	SMOTE + Tomek Links	90.8%
	CatBoost	SMOTE	90.8%

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4286

Decision Tree	SMOTE		90.7%
Gradient Boosting	SMOTE + Tomek Links		90.5%

The results reveal that **Extra Trees** paired with **SMOTE + Tomek Links** achieves the highest average accuracy of **91.9%**, ranking it as the top-performing combination. This result underscores the effectiveness of combining the ensemble strength of Extra Trees with the advanced balancing capabilities of SMOTE + Tomek Links. Similarly, **Random Forest** with SMOTE + Tomek Links secures the second position, achieving a close average accuracy of **90.8%**, further demonstrating the synergy between ensemble methods and this hybrid sampling technique.

CatBoost, a gradient boosting algorithm, ranks third with an average accuracy of **90.8%** when paired with SMOTE. While slightly lower in rank, CatBoost remains highly competitive due to its consistent performance across splits and its inherent ability to handle structured and imbalanced datasets. **Decision Tree**, another ensemble method, and **Gradient Boosting** also deliver strong results, ranking fourth and fifth, respectively, further validating the utility of SMOTE-based techniques in improving classification performance.

These rankings highlight the importance of selecting appropriate algorithm-sampling combinations to optimize model performance in imbalanced datasets. The dominance of ensemble and boosting methods, coupled with SMOTE and its variants, underscores their robustness and adaptability in diverse classification tasks. This analysis provides a clear pathway for prioritizing algorithm and sampling technique pairings in future applications.

4. DISCUSSIONS

The findings from this study highlight the critical importance of combining effective data balancing techniques with robust machine learning algorithms to address the challenges posed by imbalanced datasets. The results consistently demonstrate that **SMOTE** and its advanced variants, particularly **SMOTE + Tomek Links**, significantly enhance model performance by generating balanced training datasets. These techniques not only mitigate the adverse effects of class imbalance but also improve class separability by synthesizing minority class samples and removing noisy or overlapping data points. As a result, they enable models to achieve higher accuracy, recall, and precision, especially in scenarios with severe imbalances.

Among the algorithms evaluated, **ensemble models** such as **Extra Trees** and **Random Forest** excel in handling complex relationships within structured data. Their inherent capability to aggregate diverse decision-making processes across multiple trees makes them highly effective in learning intricate patterns, even in the presence of class imbalance. These models consistently rank at the top, particularly when paired with SMOTE + Tomek Links, achieving remarkable accuracy and balanced performance across all metrics.

In addition, **CatBoost**, a gradient boosting algorithm, emerges as a highly reliable choice for imbalanced datasets. It demonstrates competitive performance comparable to ensemble models, achieving high accuracy and robust metric scores across multiple train-test splits. CatBoost's ability to handle categorical features natively, coupled with its resistance to overfitting, makes it particularly suited for structured data classification tasks in imbalanced settings.

The results emphasize the need for a thoughtful combination of algorithm selection and data balancing techniques to maximize model performance in imbalanced datasets. While ensemble and boosting methods consistently outperform simpler models, their success depends heavily on the quality of the training data, underscoring the importance of employing advanced sampling techniques

such as SMOTE + Tomek Links. These findings provide a robust foundation for addressing class imbalance in machine learning applications and suggest clear strategies for achieving optimal classification outcomes.

5. CONCLUSION

This study highlights the critical role of pairing effective data balancing techniques with robust machine learning algorithms to address class imbalance in classification tasks. The results demonstrate that SMOTE and its advanced variant, SMOTE + Tomek Links, significantly improve model performance by generating balanced datasets and enhancing class separability, with SMOTE + Tomek Links emerging as the most effective technique. Ensemble models like Extra Trees and Random Forest consistently achieved the highest accuracy of 96.7% when paired with SMOTE-based techniques, leveraging their ability to handle complex relationships in structured data. Similarly, CatBoost proved to be a reliable boosting algorithm, delivering competitive performance even in highly imbalanced scenarios. The study also underscores the importance of train-test split ratios, with the 90:10 split yielding the best results, suggesting that larger training datasets enhance the model's learning capabilities. These findings provide a robust framework for improving classification performance in imbalanced datasets and offer a foundation for future work, including hyperparameter optimization, advanced data augmentation methods, and broader validation on diverse datasets.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

Acknowledgement is only addressed to funders or donors and object of research. Acknowledgement can also be expressed to those who helped carry out the research.

REFERENCES

[1] J. M. Johnson and T. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1-54, 2019. DOI: 10.1186/s40537-019-0192-5.

[2] L. A. Sevastyanov and E. Shchetinin, "On methods for improving the accuracy of multi-class classification on imbalanced data," *Pattern Recognition and Image Analysis*, vol. 30, no. 1, pp. 70–82, 2020. DOI: 10.14357/19922264200109.

[3] G. Du et al., "Graph-based class-imbalance learning with label enhancement," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 530–544, 2021. DOI: 10.1109/TNNLS.2021.3133262.

[4] B. S. Raghuwanshi and S. Shukla, "Class imbalance learning using underbagging-based kernelized extreme learning machine," *Neurocomputing*, vol. 329, pp. 172–187, 2019. DOI: 10.1016/j.neucom.2018.10.056.

[5] M. Bader-El-Den, E. Teitei, and T. Perry, "Biased random forest for dealing with the class imbalance problem," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 2163–2172, 2019. DOI: 10.1109/TNNLS.2018.2878400.

[6] O. Wu, "Rethinking class imbalance in machine learning," *ArXiv*, vol. abs/2305.03900, pp. 1–15, 2023. DOI: 10.48550/arXiv.2305.03900.

[7] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models," *Data Science in Finance and Economics*, vol. 4, no. 1, pp. 1–21, 2023. DOI: 10.3934/dsfe.2023021.

[8] R. Choudhary and S. Shukla, "A clustering-based ensemble of weighted kernelized extreme learning machine for class imbalance learning," *Expert Systems with Applications*, vol. 164, p. 114041, 2021. DOI: 10.1016/J.ESWA.2020.114041.

[9] S. Rezvani and X. Wang, "Class imbalance learning using fuzzy ART and intuitionistic fuzzy twin support vector machines," *Information Sciences*, vol. 578, pp. 659–682, 2021. DOI: 10.1016/J.INS.2021.07.010.
[10] E. Jiang, "A hybrid learning framework for imbalanced classification," *International Journal of Intelligent Information Technologies*, vol. 19, no. 1, pp. 1–15, 2022. DOI: 10.4018/ijiit.306967.
[11] S. Rezvani and X. Wang, "Class imbalance learning using fuzzy ART and intuitionistic fuzzy twin support vector machines," *Information Sciences*, vol. 578, pp. 659–682, 2021. DOI: 10.1016/J.INS.2021.07.010.
[12] O. Wu, "Rethinking class imbalance in machine learning," *ArXiv*, vol. abs/2305.03900, pp. 1–15, 2023. DOI: 10.48550/arXiv.2305.03900.
[13] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models," *Data Science in Finance and Economics*, vol. 4, no. 1, pp. 1–21, 2023. DOI: 10.3934/dsfe.2023021.
[14] L. Lhaura et al., "Enhancing machine learning model performance in addressing class imbalance," *CogITo Smart Journal*, vol. 10, no. 1, 2024. DOI: 10.31154/cogito.v10i1.626.478-490.
[15] S. Wang, L. L. Minku, N. Chawla, and X. Yao, "Learning from data streams and class imbalance," *Connection Science*, vol. 31, no. 1, pp. 103–104, 2019. DOI: 10.1080/09540091.2019.1572975.
[16] Z. Liu et al., "Handling inter-class and intra-class imbalance in class-imbalanced learning," *Proceedings of ICML*, 2021.
[17] S. Mirsadeghi, H. Bahsi, R. Vaarandi, and W. Inoubli, "Learning from few cyber-attacks: Addressing the class imbalance problem in machine learning-based intrusion detection in software-defined networking," *IEEE Access*, vol. 11, pp. 140428–140442, 2023. DOI: 10.1109/ACCESS.2023.3341755.
[18] J. Shetty and G. Shobha, "Handling class imbalance in Google cluster dataset using a new hybrid sampling approach," *Journal of Advances in Information Technology*, vol. 14, no. 5, pp. 934–940, 2023. DOI: 10.12720/jait.14.5.934-940.
[19] M. Abdelhamid and A. Desai, "Balancing the scales: A comprehensive study on tackling class imbalance in binary classification," *ArXiv*, vol. abs/2409.19751, 2024. DOI: 10.48550/arXiv.2409.19751.
[20] M. E. Sánchez-Gutiérrez and P. P. González-Pérez, "Addressing the class imbalance in tabular datasets from a generative adversarial network approach in supervised machine learning," *Journal of Algorithms & Computational Technology*, vol. 17, no. 1, pp. 151–168, 2023. DOI: 10.1177/17483026231215186.
[21] Z. Chen, J. Duan, L. Kang, and G. Qiu, "Class-imbalanced deep learning via a class-balanced ensemble," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5626–5640, 2021. DOI: 10.1109/TNNLS.2021.3071122.
[22] E. Rendón et al., "Density-based clustering to deal with highly imbalanced data in multi-class problems," *Mathematics*, vol. 11, no. 18, 2023. DOI: 10.3390/math11184008.
[23] J. Du, G. Qiu, Y. Lin, and S. Li, "Graph-based class-imbalance learning with label enhancement," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 530–544, 2021. DOI: 10.1109/TNNLS.2021.3133262.
[24] A. Thumpati and Y. Zhang, "Towards optimizing performance of machine learning algorithms on unbalanced dataset," *Artificial Intelligence & Applications*, vol. 13, no. 1, pp. 131–140, 2023. DOI: 10.5121/csit.2023.131914.
[25] M. Ayyannan, "Accuracy enhancement of machine learning model by handling imbalance data," *2024 International Conference on Expert Clouds and Applications (ICOECA)*, vol. 1, pp. 593–599, 2024. DOI: 10.1109/ICOECA62351.2024.00109.
[26] S. Budania, T. Kumar, H. Kumar, and G. Nikam, "Hybrid machine intelligence for imbalanced data," *Social Science Research Network*, vol. 36, no. 4, pp. 441–452, 2020. DOI: 10.2139/ssrn.3602531.

[27] H. Kaur, H. Pannu, and A. Malhi, "A systematic review on imbalanced data challenges in machine learning," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–36, 2019. DOI: 10.1145/3343440.

[28] V. H. Barella et al., "Assessing the data complexity of imbalanced datasets," *Information Sciences*, vol. 553, pp. 83–109, 2021. DOI: 10.1016/j.ins.2020.12.006.

[29] M. Zheng, F. Wang, X. Hu, Y. Miao, H. Cao, and M. Tang, "A method for analyzing the performance impact of imbalanced binary data on machine learning models," *Axioms*, vol. 11, no. 6, p. 607, 2022. DOI: 10.3390/axioms11110607.

[30] S. Ashraf and T. Ahmed, "Machine learning shrewd approach for an imbalanced dataset conversion samples," *Journal of Engineering and Technology*, vol. 11, no. 3, pp. 115–123, 2020.

[31] H. Patel, D. Rajput, O. Stan, and L. Miclea, "A new fuzzy adaptive algorithm to classify imbalanced data," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 15–29, 2022. DOI: 10.32604/cmc.2022.017114.

[32] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models," *Data Science in Finance and Economics*, vol. 4, no. 2, pp. 25–40, 2023. DOI: 10.3934/dsfe.2023021.

[33] H. Du, Y. Zhang, K. Gang, L. Zhang, and Y. Chen, "Online ensemble learning algorithm for imbalanced data stream," *Applied Soft Computing*, vol. 107, p. 107378, 2021. DOI: 10.1016/J.ASOC.2021.107378.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 627-640
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4286