A hybrid oversampling for imbalanced data using incremental SMOTE-KMeans and cluster-structured positive class-SVM

Hartono Hartono [a,1,*], Muhammad Khahfi Zuhanda [a,2], Rahmad Syah [a,3]

[a] Department of Magister Program in Informatics, Postgraduate School, Universitas Medan Area, Medan, Indonesia
[1] hartono@staff.uma.ac.id; [2] khahfi@staff.uma.ac.id; [3] rahmadsyah@uma.ac.id
* corresponding author

ARTICLE INFO

Article history
Received March 1, 2026
Revised April 4, 2026
Accepted July 29, 2026
Available online August 31, 2026

Keywords
Class imbalance
Incremental SMOTE-KMeans
Cluster-Structured Positive Class-SVM
Structure-aware classification

ABSTRACT

Class imbalance remains a significant challenge in classification tasks, particularly when the minority class exhibits complex internal structures. This study proposes a unified framework that integrates Incremental SMOTE-KMeans with a cluster-structured positive class-SVM (CS-PC-SVM) to jointly address data imbalance and structural heterogeneity. The proposed method introduces structural alignment between oversampling and classification by generating synthetic samples only within reliable clusters and modeling the minority class as multiple subgroups. This design reduces noise, preserves local data structure, and enables more adaptive decision boundaries. Experimental results on five benchmark datasets demonstrate that the proposed approach achieves consistently strong and balanced performance, with Accuracy up to 0.981, G-Mean 0.969, Precision 0.967, and Recall 0.962, outperforming or remaining competitive with existing methods. These findings highlight the effectiveness of integrating structure-guided data generation with structure-aware classification for improving robustness in imbalanced learning scenarios.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

Imbalanced data classification remains a critical challenge in machine learning [1], particularly in applications where minority instances correspond to rare but high-impact events [2]. In such cases, conventional classifiers tend to be dominated by the majority class [3], leading to poor recognition of minority patterns [4]. To alleviate this issue, data-level approaches such as Synthetic Minority Oversampling Technique (SMOTE) have been widely adopted [5], as they artificially balance the dataset by generating new minority samples through interpolation [6]. However, a fundamental limitation of SMOTE lies in its lack of structural awareness [7]. Synthetic samples are generated based solely on local neighborhoods [8], without considering whether those neighborhoods belong to reliable regions of the feature space [9]. As a consequence, SMOTE may produce samples in ambiguous areas [10], such as regions near class boundaries [11], overlapping zones [12], or noisy distributions [13]. This not only reduces the quality of the generated data but can also mislead the classifier during training [14].

To address this issue, it becomes necessary to introduce a mechanism that determines where oversampling should be performed in a principled manner [15]. In this work, incremental k-means is utilized precisely for this purpose, not as a clustering objective itself, but as a structural guide for SMOTE [16]. Specifically, incremental k-means is employed to partition the dataset into localized regions that reflect the underlying distribution of the data. Within this partitioned space, each cluster can be interpreted as a region with a certain level of reliability for generating synthetic samples [17].

Instead of allowing SMOTE to operate indiscriminately across all samples, oversampling is restricted to clusters that exhibit a sufficient presence of minority instances [18]. These clusters can be interpreted as safe regions, where interpolation between minority samples is more likely to produce meaningful and representative synthetic data. In this way, incremental k-means transforms SMOTE from a purely local interpolation method into a structure-aware data generation process [19]. This integration has two important implications. First, it significantly reduces the likelihood of generating noisy or misleading synthetic samples, as the interpolation process is confined to regions that are statistically more reliable. Second, it improves the diversity of generated samples by distributing them across multiple meaningful regions of the minority class. Nevertheless, improving data generation alone is not sufficient. A critical yet often overlooked issue is the internal heterogeneity of the minority class [20]. Most existing studies primarily focus on balancing class distribution through oversampling techniques [21], but they implicitly assume that the minority class is homogeneous [22]. Standard classifiers such as conventional SVM further reinforce this limitation by constructing a single global decision boundary that treats the minority class as one unified group [23]. This assumption is unrealistic in many real-world scenarios, leading to suboptimal performance when minority subregions are scattered [24], locally dense, or partially overlapping with the majority class [25]. To overcome this limitation, this study proposes the integration of a cluster-structured positive class-SVM (CS-PC-SVM) as the classification component. Unlike conventional classifiers, CS-PC-SVM explicitly incorporates the structural information derived from clustering, enabling the minority class to be modeled as multiple subgroups rather than a single entity. This design allows the decision function to adapt locally to different regions of the feature space, improving its ability to capture intra-class variability and handle overlapping regions. As a result, the proposed classifier provides more precise and flexible decision boundaries, particularly in complex imbalanced scenarios where traditional models tend to fail.

The integration of incremental k-means, SMOTE, and CS-PC-SVM forms a unified and gap-driven framework that jointly addresses both distributional imbalance and structural heterogeneity. Incremental k-means identifies the intrinsic subregions of the data, SMOTE leverages this structural information to perform controlled and context-aware oversampling, and CS-PC-SVM utilizes the same structure to construct an adaptive classification model. Through this coherent design, the proposed method effectively resolves the dual limitations of existing approaches, leading to improved robustness and classification performance. Despite the growing body of research on imbalanced learning, a critical gap remains in the lack of structural consistency between oversampling and classification stages. Existing oversampling methods, including clustering-based SMOTE variants, primarily focus on improving data distribution but do not explicitly align the generated samples with the classifier's decision mechanism [26]. Conversely, classifier-level approaches such as cost-sensitive SVM focus on optimizing decision boundaries but assume that the input data is already well-structured [27]. This separation limits their effectiveness, particularly when synthetic samples are generated without considering the classifier's behavior [28]. Furthermore, although the presence of subgroup structures within the minority class has been recognized, most existing methods fail to exploit this structure simultaneously during both data generation and classification [29]. To address these limitations, this study proposes a unified framework that integrates structure-guided oversampling and structure-aware classification. Incremental k-means is used to guide region-wise synthetic sample generation, while CS-PC-SVM leverages the same structural information to construct adaptive decision boundaries. The novelty of this work lies in establishing a shared structural foundation across both stages, ensuring that the augmented data is not only balanced but also aligned with the classifier.

2. Method

2.1. Handling Class Imbalance Problem

The class imbalance problem has long been recognized as a critical challenge in classification systems, as disproportionate class distributions often bias learning models toward the majority class while degrading performance on minority instances [30]. To address this challenge, a wide range of approaches have been proposed, which can broadly be categorized into data-level methods, algorithm-level methods, and hybrid frameworks [31].

At the data level, oversampling techniques have been extensively studied, with the Synthetic Minority Oversampling Technique (SMOTE) being one of the most widely adopted methods [32]. However, classical SMOTE often fails to consider local data structure and density, which may lead to noisy or overlapping synthetic samples, especially in regions with complex decision boundaries [33]. To mitigate these limitations, several SMOTE variants have been introduced, including Borderline-SMOTE [21]. While these methods improve minority representation, they still operate locally and do not explicitly incorporate structural information of the feature space [34]. To better preserve data structure, clustering-based oversampling methods have gained increasing attention. These approaches integrate clustering techniques with SMOTE to guide synthetic sample generation within more reliable regions. Cluster-based SMOTE methods aim to reduce noise by restricting oversampling to clusters with higher minority density [35]. Among these, Incremental SMOTE-KMeans has demonstrated improved performance by generating synthetic samples within selected clusters that are considered informative and representative. However, despite these improvements, clustering-based oversampling methods primarily focus on data distribution refinement and do not explicitly address how the resulting data structure should be utilized during classification [17].

At the algorithm level, Support Vector Machine (SVM) has been widely applied to imbalanced classification due to its strong generalization ability and margin-based learning. Extensions such as weighted and cost-sensitive SVM improve sensitivity to minority samples by assigning different misclassification costs. However, these approaches generally assume that the minority class is homogeneous and rely on a single global decision boundary. In practice, minority classes often exhibit internal heterogeneity, where samples are distributed across multiple subregions. To address this, structure-aware approaches, including cluster-based SVM models, have been developed to incorporate subgroup information and enable more adaptive decision boundaries [36].

In parallel, hybrid methods combining data-level and algorithm-level strategies have been explored, but most treat oversampling and classification as separate stages, leading to limited effectiveness in complex scenarios [37]. To address this gap, this study proposes a unified framework integrating Incremental SMOTE-KMeans and cluster-structured positive class-SVM (CS-PC-SVM). The method performs structure-guided oversampling within reliable regions while leveraging the same structural information for classification, enabling better alignment between data generation and decision modeling. As summarized in Table 1, this approach integrates both distributional and structural aspects, providing a more consistent and effective solution for imbalanced classification.

Table 1. Comparison of Related Works and Proposed Method

Method	Data-Level Balancing	Cluster Awareness	Classification Strategy	Main Limitations
SMOTE [38]	Yes	No	Various	Ignores local structure; may generate noisy samples
Borderline-SMOTE [21]	Yes	Partial	Various	Sensitive to noise near decision boundaries
ADASYN [34]	Yes	Partial	Various	May over-emphasize difficult or noisy samples
Cluster-Based SMOTE [35]	Yes	Yes	Various	Static clustering; limited integration with classifier
Incremental SMOTE-Kmeans [17]	Yes	Yes	Various	Focuses on data generation without classifier adaptation
SVM [36]	No	No	Margin-based	Single global boundary; weak on imbalanced data
Weighted / Cost-Sensitive SVM [39]	No	No	Margin-based	No structural awareness of minority subgroups
Proposed Method	**Yes**	**Yes**	**CS-PC-SVM**	**Integrated structure-aware learning**

2.2. Incremental SMOTE-Kmeans

Incremental SMOTE K-Means is a structure-aware oversampling approach designed to address class imbalance by considering the spatial distribution of data within the feature space. The overall stages of the proposed method are illustrated in Fig. 1, while the detailed procedure is presented in Algorithm 1 [17].

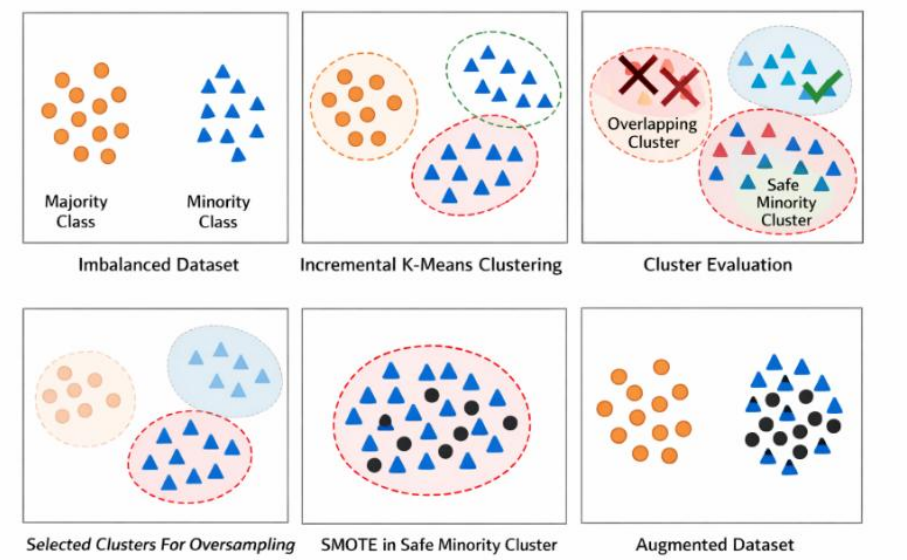


Fig. 1. Workflow of Incremental SMOTE-KMeans.

The pseudocode of Incremental SMOTE-KMeans is shown in Fig. 2 [17].

Algorithm 1 Incremental K-Means SMOTE

Input: Training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, imbalance threshold τ, total synthetic samples G
Output: Augmented dataset D^{aug}
$D^+ \leftarrow \{x_i \mid y_i = +1\}$, $D^- \leftarrow \{x_i \mid y_i = -1\}$ ▷ Separate classes
$\mathcal{F} \leftarrow \emptyset$, $\mathcal{S} \leftarrow \emptyset$
Apply Incremental K-Means on D to obtain cluster set $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$ ▷ Partition data into regions
for each cluster C_j in $\mathcal{C}$ **do**
  Compute minority count n_j^+ and majority count n_j^-
  Compute region ratio $r_j \leftarrow \dfrac{n_j^- + 1}{n_j^+ + 1}$
  if $n_j^+ > 0$ and $r_j < \tau$ **then**
    $\mathcal{F} \leftarrow \mathcal{F} \cup \{C_j\}$ ▷ Select safe cluster
  else
    Discard C_j ▷ Majority-dominated or overlapping cluster
  end if
end for
for each selected cluster C_j in $\mathcal{F}$ **do**
  Determine the number of synthetic samples g_j
  for $q = 1$ to g_j **do**
    Apply SMOTE within C_j to generate synthetic minority sample x_{new}
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x_{new}\}$
  end for
end for
Construct augmented dataset $D^{aug} \leftarrow D \cup \{(x, +1) \mid x \in \mathcal{S}\}$ ▷ Minority class increases
Return augmented dataset D^{aug}

Fig. 2. Incremental SMOTE-KMeans

The main stages of the method are described as follows. First, the dataset is partitioned into multiple localized regions using incremental clustering. This step aims to capture the intrinsic distribution of the data and enable region-wise analysis instead of treating the dataset as a single global structure. Second,

each cluster is evaluated based on its class composition. The imbalance level within each cluster is measured using a region-wise ratio.

$$r_j = \frac{n_j^- + 1}{n_j^+ + 1} \tag{1}$$

where r_j denotes the region-wise imbalance ratio, n_j^+ and n_j^- represent the number of minority and majority samples in the cluster C_j, respectively. A cluster is considered reliable if it satisfies:

$$r_j < T \tag{2}$$

where r_j denotes the region-wise imbalance ratio, T is a predefined threshold. Clusters that do not satisfy this condition are excluded from the oversampling process.

Only the selected reliable clusters are used for synthetic data generation. This ensures that the augmentation process is performed in meaningful regions and avoids introducing noise from overlapping or majority-dominated areas.

Synthetic samples are generated within the selected clusters using a controlled interpolation strategy:

$$x_{new} = x_p + \alpha(x_{nn} - x_p), \alpha \in (0,1) \tag{3}$$

where x_{new} denotes synthetic samples, α is the coefficient that controls the position of the synthetic sample, x_p is a minority sample and x_{nn} is another minority sample within the same cluster. This formulation ensures that the generated samples remain consistent with the local data structure.

Finally, the augmented dataset is constructed as:

$$D^{aug} = D \cup \{(x, +1) | x \in S\} \tag{4}$$

where D^{aug} denotes the augmented dataset, x denotes a synthetic data sample, D is an original dataset, S denotes the set of generated synthetic samples. As a result, the number of minority instances increases while preserving the structural integrity of the dataset, leading to improved learning performance.

2.3. Cluster-Structured Positive Class-SVM

Cluster-Structured Positive Class-SVM (CS-PC-SVM) is employed to incorporate subgroup information into the classification process, ensuring that the learned model remains aligned with the underlying structure revealed during earlier stages, as can be seen in Fig. 3 and the pseudocode in Algorithm 2 (Fig.4) [40].

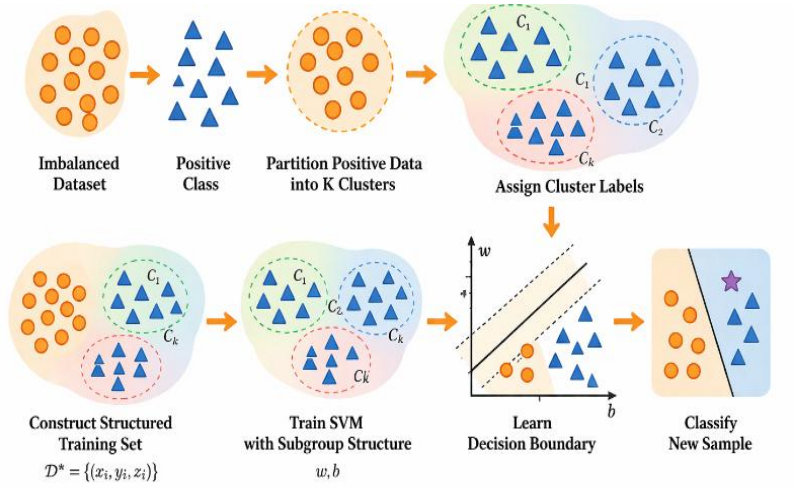


Fig. 3. Workflow of Cluster-Structured Positive Class-SVM

Algorithm 2 Cluster-Structured Positive Class-SVM (CS-PC-SVM)

Input: Training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, number of positive clusters K, penalty parameters (C^-, C_o^+, C_s^+), learning iterations T
Output: Trained CS-PC-SVM classification model
$D^+ \leftarrow \{x_i \mid y_i = +1\}$, $D^- \leftarrow \{x_i \mid y_i = -1\}$ ▷ Separate classes
$C \leftarrow \emptyset$
Partition D^+ into K clusters
for each cluster C_k in positive class **do**
    Compute centroid μ_k
    Assign cluster label $z_i \leftarrow k$ for all $x_i \in C_k$
    Add C_k to C
end for
Construct structured training set $\mathcal{D}^* = \{(x_i, y_i, z_i)\}$ ▷ Add subgroup structure
for each $(x_i, y_i, z_i) \in \mathcal{D}^*$ **do**
    if $y_i = -1$ **then**
        $w_i \leftarrow C^-$
    else if $y_i = +1$ and x_i is original **then**
        $w_i \leftarrow C_o^+$
    else
        $w_i \leftarrow C_s^+$
    end if
end for
Initialize SVM parameters: $w \leftarrow 0$, $b \leftarrow 0$
for $t = 1$ to T **do**
    for each $(x_i, y_i, z_i) \in \mathcal{D}^*$ **do**
        if $y_i(w^T x_i + b) < 1$ **then**
            $w \leftarrow w - \eta(\lambda w - w_i y_i x_i)$
            $b \leftarrow b + \eta w_i y_i$
        else
            $w \leftarrow w - \eta\lambda w$
        end if
    end for
end for
Return trained CS-PC-SVM (w, b) and classify new sample by

$$\hat{y} = \text{sign}(w^T x + b)$$

Fig. 4. Cluster-structured positive class-SVM algorithm.

In imbalanced classification problems, the minority class often exhibits internal heterogeneity, where samples are distributed across multiple subregions. To address this limitation, this study proposes a Cluster-Structured Positive Class-SVM (CS-PC-SVM), which incorporates subgroup information of the minority class into the learning process to better model its intrinsic distribution.

Given a training dataset:

$$D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathcal{R}^d, y_i \in \{-1, +1\} \tag{5}$$

Here, $D = \{(x_i, y_i)\}_{i=1}^{N}$ denotes the dataset, where $x_i \in \mathcal{R}^d$ represents the *i-th* data sample as a *d*-dimensional feature vector and $y_i \in \{-1, +1\}$ denotes its corresponding class label.
The dataset is separated into positive and negative subsets:

$$D^+ = \{x_i | y_i = +1\}, D^- = \{x_i | y_i = -1\} \tag{6}$$

Here, $D^+ = \{x_i | y_i = +1\}$ and $D^- = \{x_i | y_i = -1\}$ denote the subsets of the dataset corresponding to the positive (minority) and negative (majority) classes, respectively.

The positive class D^+ is further partitioned into K cluster.

$$D^+ = \bigcup_{k=1}^{K} C_k \tag{7}$$

Here, $D^+ = \bigcup_{k=1}^{K} C_k$ denotes that the positive class is partitioned into K clusters, Where each cluster C_k represents a subgroup of the minority class.

Each sample $x_i \in D^+$ is assigned a cluster label:

$$z_i \in \{1,2,\dots,K\} \tag{8}$$

Here, $z_i \in \{1,2,\dots,K\}$ denotes the cluster label assigned to each sample $x_i \in D^+$, indicating its membership in one of the K clusters.

This clustering process allows the model to capture the internal structure of the minority class.

To incorporate both class imbalance and subgroup structure, a weighted SVM formulation is adopted. Each sample is assigned a penalty weight w_i.

$$w_i = \begin{cases} C^-, y_i = -1, \\ C_o^+, y_i = +1 \ (Original\ Samples) \\ C_s^+ = +1 \ (Synthetic\ Samples) \end{cases} \tag{9}$$

Here, w_i denotes the penalty weight assigned to the i-th sample, y_i represents the class label of the sample (with y_i=+1 for the minority class and y_i=−1 for the majority class), and C^-, C_o^+, and C_s^+ are the penalty parameters controlling the misclassification cost for majority samples, original minority samples, and synthetic minority samples, respectively.

Subject to:

$$y_i(w^T\emptyset(x_i) + b) \geq 1 - \xi_{i,}\ \xi_{i,} \geq 0, i = 1,\dots,N \tag{10}$$

Here, w denotes the weight vector of the separating hyperplane, $\emptyset(x_i)$ represents the feature mapping of sample x_i into the transformed space, b is the bias term, and $\xi_{i,}$ is the slack variable that allows margin violations for the i-th sample.

After solving the optimization problem, the decision function is obtained as:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \tag{11}$$

where α_i are the Lagrange multipliers, $K(\cdot,\cdot)$ is the kernel function, x_i are the support vectors from the training data, y_i are their corresponding class labels, and b is the bias term.

A new sample x is classified as:

$$\hat{y} = \begin{cases} +1, f(x) \geq 0 \\ -1, f(x) < 0 \end{cases} \tag{12}$$

where $\hat{y}$ denotes the predicted class label and $f(x)$ is the decision function.

2.4. Proposed Method

The proposed method leverages incremental SMOTE-KMeans to generate reliable synthetic samples and employs CS-PC-SVM to capture the internal structure of the minority class for improved classification performance. Fig. 5 illustrates the overall workflow of the proposed method. The process begins with an incremental clustering stage, where the dataset is partitioned into localized regions to capture its intrinsic distribution. This step enables the identification of meaningful data structures rather than treating the dataset as a homogeneous space.

In the next stage, oversampling is performed in a selective manner by focusing only on reliable regions that contain sufficient minority samples. Following the data augmentation process, the classification stage employs a cluster-structured positive class-SVM (CS-PC-SVM), which incorporates subgroup information derived from clustering.

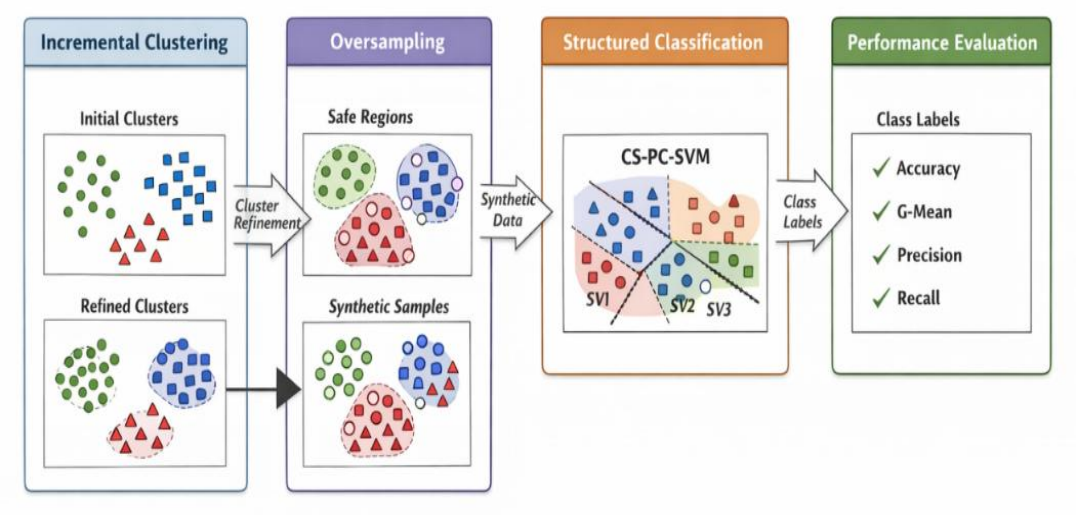


Fig. 5. Workflow of Proposed Method

The pseudocode of the proposed method is shown in Fig. 6.

Algorithm 3 Proposed Method

Require: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$
Ensure: Trained classification model
// **Stage 1: Incremental Clustering**
Partition D into clusters $C = \{C_1, C_2, \ldots, C_K\}$ using incremental K-Means
// **Stage 2: Region Selection**
for each cluster C_k **do**
  Evaluate minority proportion in C_k
  if C_k is reliable (high minority, low overlap) **then**
    Add C_k to selected clusters C^*
  end if
end for
// **Stage 3: Incremental SMOTE**
for each cluster $C_k \in C^*$ **do**
  Generate synthetic samples among minority instances in C_k
end for
Construct augmented dataset D'
// **Stage 4: Structure-Aware Classification**
Partition minority class in D' into subgroups
Train CS-PC-SVM using structured dataset D'
// **Stage 5: Prediction**
Classify new samples using trained model
return trained CS-PC-SVM model

Fig. 6. Algorithm of the proposed method

2.5. Classifier Performance

Classifier Performance will be measured using the Accuracy, G-Means, Precision, and Recall. This classifier performance measurement is carried out based on the confusion matrix which can be seen in Table 2 [41].

Table 2. Confusion Matrix

	Predictive Positive Class	Predictive Negative Class
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

The Accuracy, Accuracy, Precision, and Recall calculations can be seen in the following equation:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

$$G - Mean = \sqrt{\frac{TP}{TP+FN}} x \sqrt{\frac{TN}{TN+FP}} \quad (14)$$

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

3. Results and Discussion

3.1. Dataset

The dataset used in this study can be seen in Table 3 [41].

Table 3. Dataset

Dataset	Instances	Minority Instances	Majority Instances	Imbalance Ratio
Ecoli_0_vs_1	220	77	143	1.857
Glass1	214	76	138	1.816
Iris	150	50	100	2
Segment()	2308	329	1979	6.015
Yeast1	1484	429	1055	2.495

The datasets consist of five benchmarks with varying sizes and imbalance levels. Ecoli_0_vs_1 and Glass1 exhibit moderate imbalance, while Iris has a balanced-to-moderate ratio. Segment shows the highest imbalance, and Yeast1 presents a moderate imbalance level. These variations are important as class imbalance can significantly affect model performance, often biasing classifiers toward the majority class.

3.2. Testing for Performance

The results can be seen in Table 4-8.

Table 4. Performance for Dataset Ecoli_0_vs_1

	Proposed Method	Borderline-SMOTE	ADASYN	Cluster-Based SMOTE
Accuracy	0.977	0.967	0.923	0.969
G-Mean	0.949	0.938	0.917	0.929
Precision	0.949	0.944	0.921	0.931
Recall	0.938	0.927	0.931	0.913

The proposed method demonstrates superior and well-balanced performance across all metrics, reflecting its effectiveness in improving both overall accuracy and minority class recognition. The heatmap further confirms its consistent behavior, indicating a reliable trade-off between precision and recall for handling imbalanced datasets such as Ecoli_0_vs_1 (Fig. 7).

Table 5, the proposed method demonstrates more consistent performance across all metrics, while other methods exhibit fluctuations by performing well in certain metrics but not in others, indicating less stable handling of imbalanced data.

Table 5. Performance for Dataset Glass1

	Proposed Method	Borderline-SMOTE	ADASYN	Cluster-Based SMOTE
Accuracy	0.966	0.965	0.964	0.971
G-Mean	0.948	0.949	0.977	0.935
Precision	0.956	0.951	0.929	0.933
Recall	0.939	0.925	0.937	0.929

The performance comparison is clearly illustrated in Fig. 8, which visually highlights the consistency and effectiveness of the proposed method across all evaluation metrics.

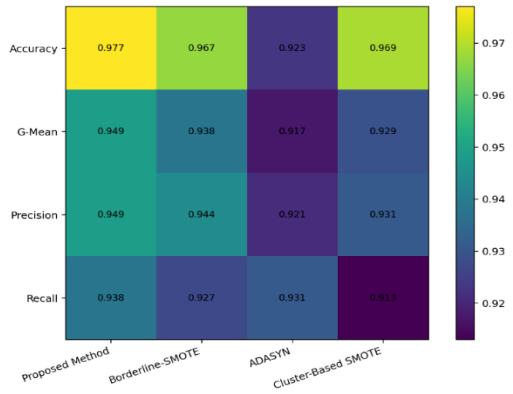


Fig. 7. Performance for Dataset Ecoli_0_vs_1

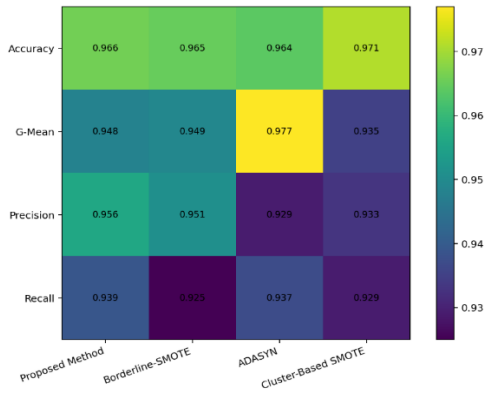


Fig. 8. Performance for Dataset Glass1

For the Iris dataset, the proposed method demonstrates a more balanced and stable performance across all evaluation metrics, indicating its ability to maintain consistency without favoring specific aspects of classification compared to the other methods (Table 6). The heatmap shows that the proposed method achieves consistently strong and well-balanced performance across all metrics on the Iris dataset (Fig. 9), demonstrating its robustness and effectiveness in handling imbalanced classification.

Table 6. Performance for Dataset Iris

	Proposed Method	Borderline-SMOTE	ADASYN	Cluster-Based SMOTE
Accuracy	0.981	0.969	0.968	0.979
G-Mean	0.969	0.961	0.965	0.967
Precision	0.967	0.957	0.955	0.959
Recall	0.961	0.954	0.949	0.954

Based on Table 7, in Dataset Segment(), the proposed method shows a more effective trade-off between accuracy and minority class detection, whereas other methods tend to emphasize certain metrics at the expense of others, indicating less balanced performance, as can be observed in the heatmap in Fig. 10, where its performance appears more uniform compared to the other methods.

Table 7. Performance for Dataset Segment()

	Proposed Method	Borderline-SMOTE	ADASYN	Cluster-Based SMOTE
Accuracy	0.957	0.949	0.948	0.958
G-Mean	0.958	0.951	0.943	0.957
Precision	0.949	0.946	0.947	0.945
Recall	0.954	0.947	0.952	0.951

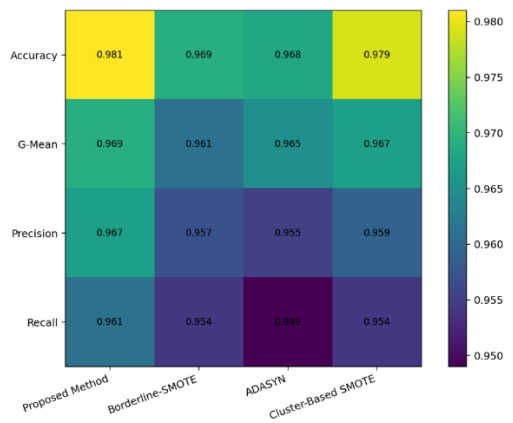


Fig. 9. Performance for Dataset Iris

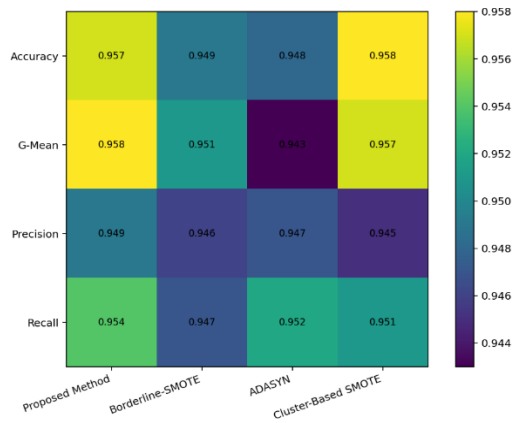


Fig. 10. Performance for Dataset Segment()

Based on Table 8, in Dataset Yeast1, the proposed method shows a more harmonious performance across all metrics, avoiding significant trade-offs observed in other methods, as can be seen in the heatmap in Fig. 11.

Table 8. Performance for Dataset Yeast1

	Proposed Method	Borderline-SMOTE	ADASYN	Cluster-Based SMOTE
Accuracy	0.969	0.951	0.952	0.961
G-Mean	0.963	0.952	0.947	0.961
Precision	0.961	0.953	0.951	0.954
Recall	0.962	0.953	0.949	0.953

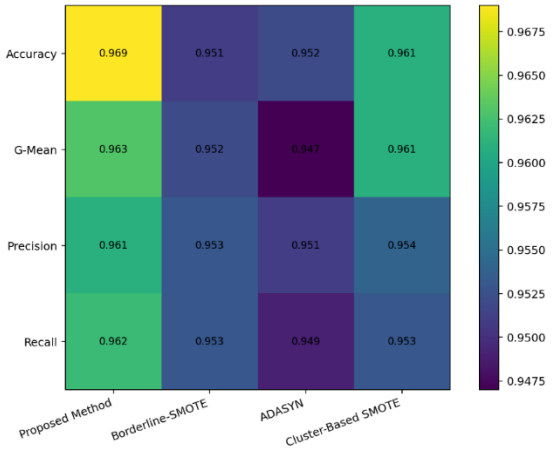


Fig. 11. Performance for Dataset Yeast1

3.3. Discussion

The experimental results across multiple benchmark datasets demonstrate that the proposed method consistently achieves competitive and, in most cases, superior performance compared to existing approaches. Beyond numerical comparisons, the heatmap visualizations in Figs. 7–11 provide clearer insight into the behavior of each method. A consistent pattern observed across these figures is that the proposed method exhibits a more uniform intensity distribution across all metrics, indicating stable and balanced performance rather than excelling in only specific aspects. This balanced behavior is mainly attributed to the structural alignment between data generation and classification. By generating synthetic samples only within reliable regions, the proposed method reduces the risk of noise and class overlap. This effect is visually evident in Fig. 7 and Fig. 11 (Ecoli_0_vs_1 and Yeast1), where the proposed method shows consistently strong responses across all metrics, while other methods display more uneven patterns, reflecting trade-offs between accuracy and minority detection.

In contrast, Fig. 8 and Fig. 10 (Glass1 and Segment) reveal a different trend. Although some methods achieve higher intensity in specific metrics (e.g., Accuracy or G-Mean), their overall patterns appear less uniform, indicating performance instability. The proposed method, while not always achieving the highest single metric, maintains a more consistent distribution across all metrics. This suggests that its advantage lies in stability rather than isolated performance gains, particularly in datasets with more overlapping or complex structures.

Despite these strengths, the proposed method depends on the effectiveness of the clustering process. Inaccurate clustering may lead to suboptimal region selection, which can affect synthetic data quality. Additionally, the integration of clustering and structured classification increases computational complexity. The heatmaps also suggest that in highly overlapping datasets, the advantage of structure-guided oversampling may be less pronounced, indicating a limitation in highly noisy environments.

Overall, both numerical results and visual patterns consistently indicate that the proposed method provides a more stable and balanced performance across different datasets. Its effectiveness is most

evident when the minority class exhibits meaningful structural patterns, making it a robust approach for imbalanced classification problems.

4. Conclusion

This study proposes a unified framework for imbalanced data classification by integrating Incremental SMOTE-KMeans with a cluster-structured positive class-SVM (CS-PC-SVM). The main contribution of this work lies in establishing a structural consistency between data-level and algorithm-level approaches. Specifically, incremental clustering is used to guide region-based oversampling, while the classification stage leverages the same structural information to model the minority class as multiple subgroups. This alignment ensures that synthetic data generation and classification are coherently integrated rather than treated as independent processes. Experimental results across multiple benchmark datasets demonstrate that the proposed method consistently provides strong and balanced performance in terms of Accuracy, G-Mean, Precision, and Recall. In particular, the method shows clear advantages in maintaining minority class representation while avoiding excessive bias toward the majority class. Despite these promising results, several limitations remain. The performance of the method is influenced by the quality of clustering, which may affect the identification of reliable regions for oversampling. In addition, the integration of multiple components introduces additional computational overhead, which may limit scalability for large-scale datasets. Future work will focus on enhancing the adaptability and efficiency of the proposed framework. Specifically, adaptive mechanisms for automatically determining the optimal number of clusters and region selection thresholds will be investigated. In addition, incorporating density-based or uncertainty-aware strategies for cluster evaluation could further improve the reliability of synthetic sample generation. Finally, exploring alternative structure-aware classifiers, such as ensemble-based or graph-based models, may further enhance performance in complex imbalanced scenarios.

Acknowledgment

The research is supported by Universitas Medan Area.

Declarations

Author contribution. Conceptualization, methodology, investigation, data curation, writing, and original draft; Muhammad Khahfi Zuhanda: supervision, resources, validation, writing, and reviewing and editing; Rahmad Syah: conceptualization, methodology, supervision, project administration, funding acquisition, writing, reviewing, and editing. All authors approved the final version
Funding statement. The research is supported by Universitas Medan Area
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] N. U. Niaz, K. M. N. Shahariar, and M. J. A. Patwary, "Class Imbalance Problems in Machine Learning: A Review of Methods and Future Challenges," in *Proceedings of the 2nd International Conference on Computing Advancements*, Mar. 2022, pp. 485–490, doi: 10.1145/3542954.3543024.

[2] J. Villanueva-García, I. Moral-Arce, and L. J. García Villalba, "A cumulative, pipeline-oriented framework for extreme imbalanced classification: Large-scale benchmarking of 4,100 rare-event models," *Expert Syst. Appl.*, vol. 312, p. 131366, May 2026, doi: 10.1016/j.eswa.2026.131366.

[3] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/s10994-022-06268-8.

[4] J. Jiang *et al.*, "A review of machine learning methods for imbalanced data challenges in chemistry," *Chem. Sci.*, vol. 16, no. 18, pp. 7637–7658, 2025, doi: 10.1039/D5SC00270B.

[5] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," *Inf. Softw. Technol.*, vol. 139, p. 106662, Nov. 2021, doi: 10.1016/j.infsof.2021.106662.

[6] L. Hu and Y. Zhang, "GDSMOTE: A Novel Synthetic Oversampling Method for High-Dimensional Imbalanced Financial Data," *Mathematics*, vol. 12, no. 24, p. 4036, Dec. 2024, doi: 10.3390/math12244036.

[7] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Min.*, vol. 16, no. 1, p. 15, Apr. 2023, doi: 10.1186/s13040-023-00330-4.

[8] A. F. Kalay, "Generating Synthetic Data With Locally Estimated Distributions for Disclosure Control," *Aust. N. Z. J. Stat.*, vol. 68, no. 1, p. e70032, Mar. 2026, doi: 10.1111/anzs.70032.

[9] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, "Challenges and limitations of synthetic minority oversampling techniques in machine learning," *World J. Methodol.*, vol. 13, no. 5, pp. 373–378, Dec. 2023, doi: 10.5662/wjm.v13.i5.373.

[10] H. Nizam-Ozogur and Z. Orman, "A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data," *Expert Syst.*, vol. 41, no. 8, pp. 1–22, Aug. 2024, doi: 10.1111/exsy.13596.

[11] C. Bunkhumpornpat, E. Boonchieng, V. Chouvatut, and D. Lipsky, "FLEX-SMOTE: Synthetic over-sampling technique that flexibly adjusts to different minority class distributions," *Patterns*, vol. 5, no. 11, p. 101073, Nov. 2024, doi: 10.1016/j.patter.2024.101073.

[12] X. Lu, X. Ye, and Y. Cheng, "An overlapping minimization-based over-sampling algorithm for binary imbalanced classification," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108107, Jul. 2024, doi: 10.1016/j.engappai.2024.108107.

[13] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, p. 37, Jan. 2025, doi: 10.3390/a18010037.

[14] S. Song, S. Yang, and M. Sun, "CS-SMOTE: An Improved Oversampling Method Combining SMOTE Method and Symmetrical Cube Scoring Mechanism," *Symmetry (Basel).*, vol. 17, no. 11, p. 1941, Nov. 2025, doi: 10.3390/sym17111941.

[15] L. Tao, H. Li, F. Wang, M. Liu, Z. Tang, and Q. Wang, "An Adaptive Safe-Region Diversity Oversampling Algorithm for Imbalanced Classification," *IEEE Access*, vol. 12, pp. 63713–63724, 2024, doi: 10.1109/ACCESS.2024.3396155.

[16] A. Abernathy and M. E. Celebi, "The incremental online k-means clustering algorithm and its application to color quantization," *Expert Syst. Appl.*, vol. 207, p. 117927, Nov. 2022, doi: 10.1016/j.eswa.2022.117927.

[17] D. S. Turan and B. Ordin, "The incremental SMOTE: A new approach based on the incremental k-means algorithm for solving imbalanced data set problem," *Inf. Sci. (Ny).*, vol. 711, p. 122103, Sep. 2025, doi: 10.1016/j.ins.2025.122103.

[18] A. Cabras, P. Ortu, T. Pisanu, P. Maxia, and R. Caocci, "Incremental Clustering for Predictive Maintenance in Cryogenics for Radio Astronomy," *Sensors*, vol. 24, no. 7, p. 2278, Apr. 2024, doi: 10.3390/s24072278.

[19] A. Zhu, Z. Hua, Y. Shi, Y. Tang, and L. Miao, "An Improved K-Means Algorithm Based on Evidence Distance," *Entropy*, vol. 23, no. 11, p. 1550, Nov. 2021, doi: 10.3390/e23111550.

[20] K. R. Mahmudah, F. Indriani, Y. Takemori-Sakai, Y. Iwata, T. Wada, and K. Satou, "Classification of Imbalanced Data Represented as Binary Features," *Appl. Sci.*, vol. 11, no. 17, p. 7825, Aug. 2021, doi: 10.3390/app11177825.

[21] M. G. Malhat, A. M. Elsobky, A. E. Keshk, H. A. Abdallah, and M. Hussein, "An approach for handling imbalanced datasets using borderline shifting," *Sci. Rep.*, vol. 16, no. 1, p. 8264, Mar. 2026, doi: 10.1038/s41598-026-39118-x.

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

[23] J. Klikowski and M. Woźniak, "Employing One-Class SVM Classifier Ensemble for Imbalanced Data Stream Classification," 2020, pp. 117–127, doi: 10.1007/978-3-030-50423-6_9.

[24] H. Najafi, G. L. Lagerwall, J. Obeysekera, and J. Liu, "Machine Learning in Climate Downscaling: A Critical Review of Methodologies, Physical Consistency, and Operational Applications," *Water*, vol. 18, no. 2, p. 271, Jan. 2026, doi: 10.3390/w18020271.

[25] P. Vuttipittayamongkol and E. Elyan, "Overlap-Based Undersampling Method for Classification of Imbalanced Medical Datasets," 2020, pp. 358–369, doi: 10.1007/978-3-030-49186-4_30.

[26] Y. Tao, Y. Zhang, and B. Jiang, "DBCSMOTE: a clustering-based oversampling technique for data-imbalanced warfarin dose prediction," *BMC Med. Genomics*, vol. 13, no. S10, p. 152, Oct. 2020, doi: 10.1186/s12920-020-00781-2.

[27] P. Cao, D. Zhao, and O. Zaiane, "An Optimized Cost-Sensitive SVM for Imbalanced Data Learning," 2013, pp. 280–292, doi: 10.1007/978-3-642-37456-2_24.

[28] A. X. Wang, V.-T. Le, H. N. Trung, and B. P. Nguyen, "Addressing imbalance in health data: Synthetic minority oversampling using deep learning," *Comput. Biol. Med.*, vol. 188, p. 109830, Apr. 2025, doi: 10.1016/j.compbiomed.2025.109830.

[29] C. Shyalika, R. Wickramarachchi, and A. P. Sheth, "A Comprehensive Survey on Rare Event Prediction," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–39, Mar. 2025, doi: 10.1145/3699955.

[30] P. Soltanzadeh, M. R. Feizi-Derakhshi, and M. Hashemzadeh, "Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach," *Pattern Recognit.*, vol. 143, p. 109721, Nov. 2023, doi: 10.1016/j.patcog.2023.109721.

[31] J. Huang, Y. Wang, and M. Wang, "Class-Adaptive Weighted Broad Learning System with Hybrid Memory Retention for Online Imbalanced Classification," *Electronics*, vol. 14, no. 17, p. 3562, Sep. 2025, doi: 10.3390/electronics14173562.

[32] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, "An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space," *Sci. Rep.*, vol. 15, no. 1, p. 23521, Jul. 2025, doi: 10.1038/s41598-025-09506-w.

[33] J. Lyu, J. Yang, Z. Su, and Z. Zhu, "LD-SMOTE: A Novel Local Density Estimation-Based Oversampling Method for Imbalanced Datasets," *Symmetry (Basel).*, vol. 17, no. 2, p. 160, Jan. 2025, doi: 10.3390/sym17020160.

[34] Y. Pristyanto, A. F. Nugraha, A. Dahlan, L. A. Wirasakti, A. Ahmad Zein, and I. Pratama, "Multiclass Imbalanced Handling using ADASYN Oversampling and Stacking Algorithm," in *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Jan. 2022, pp. 1–5, doi: 10.1109/IMCOM53663.2022.9721632.

[35] A. R. Salehi and M. Khedmati, "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data," *Sci. Rep.*, vol. 14, no. 1, p. 5152, Mar. 2024, doi: 10.1038/s41598-024-55598-1.

[36] C. Jimenez-Castaño, A. Álvarez-Meza, D. Cárdenas-Peña, A. Orozco-Gutíerrez, and J. Guerrero-Erazo, "Kreĭn twin support vector machines for imbalanced data classification," *Pattern Recognit. Lett.*, vol. 182, pp. 39–45, Jun. 2024, doi: 10.1016/j.patrec.2024.03.017.

[37] A. Salehi and M. Khedmati, "Hybrid clustering strategies for effective oversampling and undersampling in multiclass classification," *Sci. Rep.*, vol. 15, no. 1, p. 3460, Jan. 2025, doi: 10.1038/s41598-024-84786-2.

[38] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, Nov. 2021, pp. 1–8, doi: 10.1109/ICIC54025.2021.9632912.

[39] Z. Hengyu, "A New Cost-sensitive SVM Algorithm for Imbalanced Dataset," in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Jan. 2021, pp. 402–407, doi: 10.1109/ICCECE51280.2021.9342072.

[40] J. Mun, S. Bang, and J. Kim, "Weighted support vector machine for extremely imbalanced data," *Comput. Stat. Data Anal.*, vol. 203, p. 108078, Mar. 2025, doi: 10.1016/j.csda.2024.108078.

[41] J. Guo, H. Wu, X. Chen, and W. Lin, "Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification," *Appl. Soft Comput.*, vol. 150, p. 110986, Jan. 2024, doi: 10.1016/j.asoc.2023.110986.