A Hybrid Mutual-Information and Heatmap Driven Under-Sampling Algorithm for Imbalanced Binary Classification

Sehar Gul[1], Syahid Anuar[2], Hazlifah Mohd Rusli[3], Azri Azmi[4], Sadaquat Ali Ruk[5]
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia[1,2,3,4,5]
Department of Computer Science, Sukkur IBA University, 65200 Sukkur, Pakistan[1]
Department of Computer Science, Shah Abdul Latif University Khairpur, Ghotki Campus, Pakistan[5]

***Abstract*—Class imbalance is a common problem that occurs in classification where one class has much more instances than the other class. Class imbalance is especially challenging in high-stakes fields like medical diagnosis, fraud detection and predictive maintenance among others. In cases of imbalanced class distribution, models perform well while predicting the majority class but are unable to predict the minority class, which is usually more important. This paper introduces the MI-Heat (Mutual-Information and Heatmap Driven Under-sampling), a hybrid algorithm which targets the binary classification problem. The algorithm is a data-level method that combines Mutual Information (MI) for identifying important features and K-Means clustering for identification of the most important majority class samples. In addition, a distance heatmap is used to project proximities among samples and cluster centers, guiding what majority instances to retain and what to discard. Together, Mutual Information, clustering, and heatmap preserve diversity and suppress noise in order to increase the ability of the model to represent both classes equally with clarity. Performance of the MI-Heat algorithm is tested on 23 benchmark datasets and the results are seen as an improvement in classification accuracy, minority class recall and model generalization. When compared to the traditional under-sampling approaches, MI-Heat performance is consistently better, which clearly demonstrates its dominance over dealing with the class imbalance issue.**

***Keywords*—Mutual information; heat-map visualization; clustering; under-sampling; data-level; hybrid**

I. Introduction

Class imbalance refers to a prevalent situation in binary class problems, where one class heavily dominates the other, resulting in an imbalanced data distribution [1]. In these setups, machine learning systems use supervised or unsupervised learning methods to tag unstructured, typically subjective data [2]. Such asymmetry is naturally challenging in numerous domains especially in high-risk domains like medical diagnosis [3], [4], fraud detection [5], [6] and predictive maintenance. When there is class imbalance, models bias towards the majority class with resulting negative effects on their prediction performance for the minority class, typically of greater interest [7], [8], [9].

In diagnostic medicine, for instance, the influence of class imbalance may negate the ability to identify rare diseases that must be identified early if they are to be effectively treated [10]. Models learned from unbalanced datasets may be less sensitive to the minority class and consequently more likely to produce false negatives [11], [12]. These errors are particularly undesirable in medical scenarios where diseases like cancers or rare diseases must be diagnosed early [13]. [9] have found a study indicating that class imbalance has a tendency to significantly impact the performance of disease predictive models, an indicator of the significance of methods effective at dealing with differences in class distribution.

The effects of class imbalance also trickle down to fraud detection systems [14]. In the system, fraudulent transactions are the minority of the total transactions, so classifiers that have been trained using biased data find it hard to detect such critical instances. In line with [15], the monetary implications of having fraud activity misclassified are disastrous, thereby requiring it all the more to have high-precision and recall models for fraud detection. Similarly, in predictive maintenance, the strength of models relies on their ability to effectively identify failure events; thus, class imbalance resolution is paramount to prevent costly downtime and inefficiency in industrial setups [16]. To reduce the data imbalance, one way is to manipulate the dataset through sampling [17], [18], [19].The other focuses on manipulating the algorithm to reduce the effect [20]. Data-level methods target to reduce majority class instances or adding new information to the minority class. These methods involve rebalancing of the dataset using data resampling, under-sampling and over-sampling techniques [21], [22].

Although many solutions at the level of sampling and algorithms have been proposed in order to handle class imbalance, very few methods explicitly leverage structural patterns hidden within the data. Most of the under-sampling techniques based on hybrid and clustering approaches rely on distance metrics or estimations of density which do not consider how informative features interact between minority and majority classes. Consequently, such methods cannot preserve critical instance relationships or sometimes eliminate samples with discriminative information. Mutual information-based approaches represent one promising direction, yet, unfortunately, are rarely combined with spatial or structural insights such as heatmap visualization of feature relevance. An approach that combines mutual information with heatmap-driven visualization in guiding informed under-sampling decisions is highly lacking in current literature. More precisely, there is a lack in the study of techniques that find redundant majority regions by using information-theoretic measures, which also preserves meaningful minority clusters, resulting in consistent generalization in datasets under different imbalance severity

conditions. This opens a gap in the development of an MI-Heatmap-based under-sampling method that can intelligently select samples by exploiting feature-relevance patterns and data geometry instead of relying on randomness.

The remainder of this paper is organized according to the following plan. Section II introduces background and related work, including under-sampling techniques, clustering for addressing class imbalance, Mutual Information (MI) feature selection, and interpretability through heatmaps. Section III introduces the MI-Heat algorithm, explains its workflow, and its theoretical rationales, e.g., cluster separation, heatmap interpretability, and time complexity.Section IV describes the experimental setup: baselines, approaches, choice of classifier, and measures. Section V discusses and reports the results. Section VI contains the conclusion and future directions.

II. Related Work

This paper introduces a hybrid mutual information-based clustering approach which utilizes heatmap visualization, for further improvement of the process of reducing the data from majority class without losing important information. This section includes under-sampling approaches, clustering in class imbalance, mutual information and heatmap visualization for interpretability

A. Under-Sampling Approaches

To overcome the issues of class imbalance, various methods are employed like over-sampling, under-sampling and advanced ensemble techniques. For instance, methods such as SMOTE (Synthetic Minority Over-sampling Technique) enhance the representation of the minority class using synthetic sampling, improving the performance of the model [23]. Moreover, research identifies that various approaches are of varying effectiveness depending on the specific characteristics of the compared data sets, pointing to tailored solutions rather than one approach to all [24].

Random Under-sampling (RUS) is among the simplest techniques to address class imbalance. It chooses instances randomly and removes it from the majority class until both classes have equal proportions. Although this can improve a model's ability to recognize minority instances, it has the effect of inducing loss of vital information if key examples are being removed [15]. Tomek Links builds upon this by cleaning the decision boundary removing overly close majority samples to minority ones [25], [26]. Tomek Links can still eliminate useful samples accidentally and cause overfitting or misrepresent the class distribution [24]. NearMiss tries to select representative majority samples close to minority instances, retaining valuable context and increased sensitivity [27]. But it may risk losing important data, limiting generalization [28]. DBSCAN-based under-sampling, however, clumps data and removes majority points of sparse regions but keeps denser ones. While better organized, it is also dependent on tuning parameters and still risks discarding useful or noisy samples [29].

Although useful, these methods have vulnerabilities that come with them by nature: they can discard valuable majority information, reduce diversity and cause under- or over-fitting. Each approach also experiences different performance on different datasets, showing that under-sampling also lacks the more adaptable and informative strategies it needs [24]. Traditional under-sampling methods for imbalance classes, such as Random Under-sampling (RUS), Tomek Links and Near-Miss, suffer from the disadvantage of removing informative majority class samples and consequently losing useful information [30] and inducing underfitting. They can also reduce diversity in the majority class, which is a critical factor in successful decision boundary learning.

Condensed Nearest Neighbor (CNN) reduces the training set without losing the most important patterns, but selecting the optimal subset is NP-hard and most often relies on heuristics [31]. Edited Nearest Neighbor (ENN) improves classification by deleting majority samples incorrectly classified by their neighbors but may delete useful points as well [32]. As a result, these methods must balance eliminating redundancy with the preservation of informative data [33].

Even though SMOTE and other oversampling methods are useful in expanding minority class samples, they introduce noise and tend to enhance overfitting when the minority class is small.

However, under-sampling methods may results in an information loss due to extreme reduction in the majority class [34]. This will ultimately lead to models that overfit to the remaining data, particularly when the samples retained do not adequately represent the original dataset's diversity [35].

B. Clustering in Class Imbalance

Clustering is a good way to preserve structure and diversity of the majority class under under-sampling. Random or distance-based only methods treat all majority samples the same way and have a tendency to remove useful points. Clustering, specifically K-Means, first groups similar majority instances together into meaningful clusters then keep some representative samples from each cluster. This preserves important patterns and removes redundant or noisy points. There is evidence to support this: clustering can concentrate a model's attention on the minority class more strongly [36] and when combined with learned samplers like MLPs, can provide improved accuracy in difficult settings like bioinformatics [37]. DBSCAN can also help by eliminating sparse, less-informative majority points and preserving dense structure, but is hyperparameter-sensitive [29]. However, K-Means is generally preferred for its simplicity and stability it selects centroid or near-centroid samples, which improves generalization and bounds information loss [24].

C. Mutual Information (MI) in Feature Selection

Mutual information (MI) chooses the most relevant features to predict the class. By selecting the most relevant features, MI increases the ability of the model to discriminate between minority and majority classes, particularly when there is overlapping in class distribution [38], [39], [40]. Using MI before under-sampling preserves the signal and removes the clutter, so the model learns more from fewer, cleaner examples. For instance, Chen et al. demonstrate the usefulness of MI for feature selection when they indicate that it helps to identify features that have high information gain about the minority class, which can lead to improved discrimination

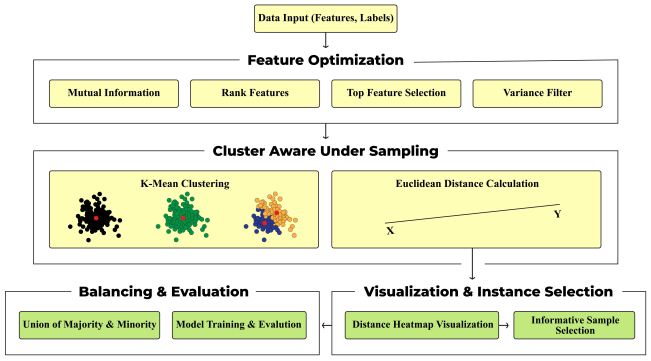


Fig. 1. Conceptual framework.

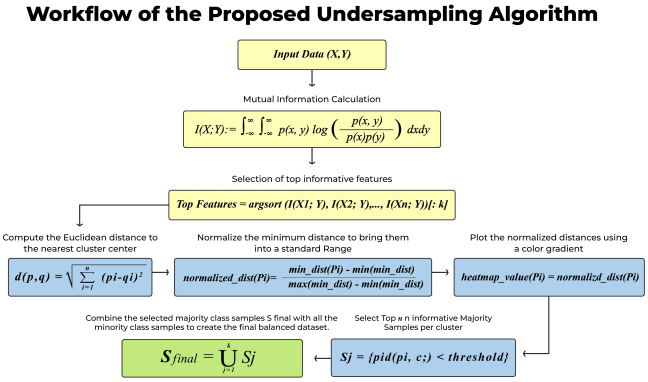


Fig. 2. Workflow of algorithm.

between the classes [41]. This is crucial because minority classes have a tendency of lacking representative data and therefore, classifiers fail to learn their characteristics. The application of MI allows one to select features that optimize the model's ability to separate minority and majority classes, thus enhancing overall classification performance, particularly for overlapping distributions [42]. Furthermore, [43] validate the applicability of MI-based feature selection in the context of imbalanced datasets, showing how it allows for the identification of pertinent features and the removal of redundant or non-informative features.

Combining clustering with MI makes it even stronger. Clustering can tell us which features to keep by eliciting structure in the data and MI ranks those features by usefulness. Adaptive, MI-based selection has also been shown to remove redundancy and keep only the most informative features [44]. Hybrid designs that couple under-sampling with heuristic feature selection also produces more quality data and results [45]. Furthermore, MI weighting models can enhance performance in high-stakes fields (e.g., medicine) by better encoding true class structure [46].

D. Heatmap Visualization for Interpretability

Heatmaps facilitate easier visualization about under-sampling choice. Distance heatmap shows nearest distance of each sample to minority points and to cluster centers, thus representative majority examples can be retained while deleting redundant/noisy ones. This is helpful to avoid class imbalance biased classifiers. Past research validates informed sampling: [47] highlight accurate sampling during training; [48] illustrate density-aware approaches preserving essential minority information; and [49] indicate heatmaps as an easy way of visualizing closeness and informing under-sampling. In practice, heatmaps facilitate practitioners to see class structure, to prioritize boundary cases and to justify what is kept or thrown away with a single, understandable figure.

However, most of these sampling methods still have key limitations. For instance, random under-sampling results in the removal of informative majority instances, introducing information loss. Most of the existing hybrid or clustering-based methods improve balance but are mostly not interpretable and cannot explain why certain samples are retained or rejected. Visualization-aided sampling methods exist, but most provide coarse or domain-specific representations that do not clearly expose feature relevance or cluster structure. Consequentially, decision transparency of current imbalance handling techniques is still limited, leaving room for development at the intersection of information-theoretic feature selection with interpretable, structure-aware sampling. Based on such shortcomings, the MI-Heat algorithm proposes a more structured and explainable alternative: Hybrid Mutual-Information and Heatmap-Driven Under-sampling.

By combining these methods, the MI-Heat (Hybrid Mutual-Information and Heatmap Driven Under-sampling) algorithm provides a new solution to class imbalance problem. By employing clustering to efficiently distribute the majority class samples, Mutual Information (MI) further helps in selecting features that are most representative in nature, making the model more sensitive to the minority class. Heatmap visualization included in the approach also makes it more interpretable, such that under-sampling itself becomes efficient as well as clear. This under-sampling hybrid approach reduces majority class representation without sacrificing the diversity and integrity of the majority class, circumventing the limitations of traditional under-sampling methods that often result in data loss and overfitting. While successful in these data-level methods, class overlap and distribution differences between the classes remain challenges. Synthetic data generation can even introduce noise or omit the actual distribution of the minority class. Furthermore, where minority samples are few, it is impossible to fully represent differences between the classes and the utility of data-level methods is limited.

III. Proposed Methodology: MI-Heat Algorithm

The Hybrid Mutual-Information and Heatmap Driven Under-sampling (MI-Heat) method starts with mutual choice of the most informative features from the data by way of Mutual Information (MI), as shown in Fig. 1. This is to effectively eliminate only the features that contribute minimal value towards classification, while keeping the features of maximum importance. Then, any feature with small variability is removed by a Variance Filter, so that relevant information can be used for further analysis. Second, the most common class instances are clustered collectively with the aid of K-Means Clustering, splitting the data into more manageable

smaller sets that are closer in nature. This assists the algorithm in achieving a better understanding of the overall structure of the majority class. Post-clustering, the algorithm computes distances from every data point to the center of the respective cluster. These are rendered visible with a heatmap, from which it is easy to spot the data points nearer to the cluster centroids and thus more significant.

Finally, the algorithm selects representative points from every cluster and, in this instance, the nearest to the cluster center. These representative points are added to all instances of the majority class and create a balanced dataset. By the proper selection of representative points from every cluster and the preservation of all majority and minority class instances, this method avoids the end dataset to be deprived of vital information regarding the data distribution and hence it becomes a balanced and informative dataset that results in improved model performance and more accuracy in prediction.The mathematical formulations utilized in the proposed algorithm 1 are elaborated in the subsequent overview section.

A. Overview

As shown in Fig. 2, the proposed approach, follows a formal workflow starting from mutual information (MI) computation and top feature selection. Clustering is performed on the features, after which heatmap is used to visualize by using the euclidean distance.

- Input data: Begin with an imbalanced dataset with feature matrix $\mathbf{X}$ and labels $\mathbf{y}$.
- Mutual information (MI) feature selection: Compute MI between each feature and the target, then keep the top k features:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\,\log\!\left(\frac{p(x,y)}{p(x)p(y)}\right). \quad (1)$$

- Variance filter: Remove near-constant features using a variance threshold to retain informative variables.
- Clustering of majority-class samples: Apply k-means to the majority-class subset to form clusters based on feature similarity.
- Euclidean distances: For majority sample $\mathbf{p}_i$ and cluster center $\mathbf{c}_j$,

$$d(\mathbf{p}_i,\mathbf{c}_j) = \sqrt{\sum_{m=1}^{n}(p_{im}-c_{jm})^2}, \quad (2)$$

  where n is the number of retained features.
- Distance heatmap generation: Compute normalized distances between majority samples and their assigned centroids and visualize them with a heatmap (darker indicates closer).
- Select informative majority samples from clusters: From each cluster, choose the n closest samples to its centroid (by sorted $d(\cdot,\cdot)$) to form S_{final}.
- Balanced dataset creation: Combine selected majority samples with all minority samples:

Algorithm 1 Hybrid Mutual-Information and Heatmap Driven Under-Sampling (MI-Heat)

Require: Training data $X_{\text{train}} \in \mathbb{R}^{n\times d}$, labels y_{train}, majority class label c_M, under-sampling ratio $r \in (0,1]$, number of top features k_f, random seed s.
Ensure: Under-sampled dataset (X', y') with balanced class distribution.

Identify class indices
$\mathcal{I}_M \leftarrow \{\, i \mid y_{\text{train}}[i] = c_M \,\}, \quad \mathcal{I}_m \leftarrow \{\, i \mid y_{\text{train}}[i] \neq c_M \,\}.$
if $|\mathcal{I}_M| = 0$ **then**
    error: "No majority instances found."
end if
Select informative and stable features
Compute mutual information scores $MI_j = MI(X_j, y_{\text{train}})$.
Retain top-k_f features with highest MI_j.
Apply variance filtering ($\sigma^2 < 10^{-5}$) to remove near-constant predictors. Resulting matrix: X_s.
Extract majority subset $X_M \leftarrow X_s[\mathcal{I}_M]$.
Determine clustering granularity $k \leftarrow \max\big(1, \lfloor |\mathcal{I}_M| \times r \rfloor\big)$.
Perform cluster analysis
Apply K-Means (k clusters, seed s) on X_M to obtain cluster centers $C = \{c_1, \ldots, c_k\}$ and labels L.
Compute proximity metrics and visualize heatmap
Compute Euclidean distances $D_{ij} = \|X_M[i] - c_j\|_2$.
Normalize the minimum distance per sample:

$$h_i = \frac{\min_j D_{ij} - \min(h)}{\max(h) - \min(h)}.$$

Visualize $\{h_i\}$ as a one-column heatmap to show density and dispersion of majority points.
Cluster-aware instance selection
for $c \in \{1, \ldots, k\}$ **do**
    $S_c = \{\, i \mid L[i] = c \,\}$.
    $n_c = \max\big(1, \lfloor |S_c| \times r \rfloor\big)$.
    Sort S_c by $D_{i,c}$ ascending.
    Retain first n_c indices to obtain S'_c.
end for
Concatenate all retained indices: $\mathcal{I}'_M = \bigcup_c S'_c$.
Form the final balanced subset
$\mathcal{I}' = \mathcal{I}'_M \cup \mathcal{I}_m; \quad X' = X_{\text{train}}[\mathcal{I}'], \quad y' = y_{\text{train}}[\mathcal{I}']$.
Reset indices for consistency.
Return (X', y').

$$\mathcal{D}_{\text{balanced}} = S_{\text{final}} \cup \mathcal{D}_{\text{minority}}. \quad (3)$$

Notation: $p(\cdot)$ denotes probability mass/density; bold symbols are vectors; $\log$ is the natural logarithm.

B. Theoretical Rationale

MI-Heat combines feature selection and clustering to address the imbalanced data issue in a more interpretable and structured manner. It combines feature relevance with under-sampling that is cluster-aware to preserve valuable patterns within the majority class without letting the minority class be overlooked. The hybrid approach has several theoretical

advantages to enable the model to learn better from both classes.

1) Relevance enhances cluster separation: In the MI-Heat model, the process begins by selecting the most informative features according to Mutual Information (MI), by using the formula discussed in Eq. (1). This process ensures that only features most effective at distinguishing between classes are retained. Operating with these features, the algorithm enhances the separation between clusters of the majority class. As a result, the clusters obtained in the subsequent K-Means clustering step are more pronounced, with better defined borders, making it easier to distinguish between the majority and minority classes. Cluster diversity avoids loss of minority-adjacent samples. Instead of simply random under-sampling the majority class samples, the MI-Heat method uses K-Means clustering to divide the majority class into smaller clusters. In this manner, diversity among the majority class is preserved. The algorithm ensures that notable, divergent minority instances near the minority class are preserved by selecting the most characteristic points from every cluster. This reduces the risk of losing valuable minority-adjacent samples, which would harm the ability of the model to learn the minority class characteristics. To construct the final balanced dataset, the representative majority samples selected from all clusters are combined with the complete minority set, as represented in Eq. (3).

2) Heatmap provides interpretable cluster-density mapping: The distance heatmap generated during the clustering phase is computed using the Euclidean metric in Eq. (2), which serve as an interpretation aid employed in helping us comprehend the structure of the majority class. The heatmap shows the distances of the majority instances to their respective cluster centers, with the intensity of the colour representing the sample density in each region. The visualization serves to aid understanding of where the data points are centered and the quality of the definition of the clusters. It gives interpretability to the procedure such that it's possible to see the spatial location of majority class points and how the algorithm decides what points to retain.

3) Time complexity analysis $\mathcal{O}(n_M k d' t)$: The time complexity of the MI-Heat algorithm is different based on the factors discussed below. n_M represents the number of majority class instances, k represents the number of clusters derived in the K-Means clustering and d' is the number of features yielded after Mutual Information and Variance Filtering. The most prominent time complexity is the K-Means clustering with a time complexity of $\mathcal{O}(n_M k d' t)$, where t represents the number of iterations the K-Means algorithm runs through. This is a sign of the clustering cost of the most frequent class data and calculating the distances between every sample and its nearest cluster center.

IV. Experimental Setup

This section reports experimental setup used to evaluate the performance of the Hybrid Mutual-Information and Heatmap Driven Under-sampling (MI-Heat) algorithm. The evaluation is done on various benchmark datasets, baseline comparison, classifier training and assessment using variety of evaluation metrics.

A. Datasets

For testing, 23 standard benchmark datasets are retrieved from UCI and KEEL in various domains such as medical, biological and industrial data. The data was split into 30% for training and 70% for testing. Out of the 30% training data, the initial count of the majority class samples is identified from the entire dataset.However, to address the class imbalance problem, the majority class is under-sampled, so that there is balance between the majority class and the minority class in the training set. This is where limited number of best majority samples were selected representing the entire majority class. These datasets were selected to be representative of a variety of imbalanced classification problems. Table I show a list of the datasets used in the experiments.

B. Baseline Methods

The performance of the MI-Heat algorithm is compared with several baseline under-sampling techniques, commonly used in imbalanced data scenarios. These include:

1) Random Under-sampling (RUS): A basic under-sampling technique that randomly selects a subset of the majority class samples, discarding the rest.

2) Tomek links: A cleaning method that removes pairs of samples from different classes that are closest to each other.

3) NearMiss: A family of algorithms that selects majority class samples closest to minority class samples based on a distance metric.

4) DBSCAN-based under-sampling: A density-based under-sampling method that uses DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to find dense regions of the majority class and reduce instances accordingly.

5) ENN (Edited Nearest Neighbors): An algorithm that removes samples from the majority class that are misclassified by their nearest neighbors.

6) CNN (Condensed Nearest Neighbors): A similar technique to ENN but instead condenses the dataset by retaining only the necessary majority class samples that are not misclassified.

C. Choice of Classifier

To evaluate the effectiveness of the proposed method, Random Forest(RF) is selected as a classifier. Random Forest is a robust and efficient ensemble learning method that is well-known for its ability to handle imbalanced data and reduce overfitting. With under-sampling, it further reduces majority-class bias and improves minority detection [50]. Empirical data prove RF exhibits consistency across data-acquisition environments and various types of datasets. In medical science, RF has worked better than traditional classifiers, i.e., gestational diabetes prediction, proving its actual-world use [51]. It is also good in complex relations under different rebalancing methods [52]. It is trained on the under-sampled data of the MI-Heat algorithm as well as the baseline methods. The Random Forest classifier was implemented using the scikit-learn library with the following parameters:

- Number of trees : 300

TABLE I. SUMMARY OF THE DATASETS USED IN THE EXPERIMENTS

Dataset	Samples	Features	Minority Ratio	Actual Majority	Selected Majority	Source
ECOLI-0_Vs_1	336	7	1:3	56	31	UCI
Spambase	4601	57	1:1	1984	1246	UCI
abalone20-vs-8-9-10	1333	8	1:9	17	9	UCI
abalone9-18	731	8	1:18	32	20	UCI
cardiotocography	1349	21	1:5	329	190	KEEL
ecoli-0-1-3-7_vs_2-6	327	7	1:2	4	2	UCI
ecoli-0-1-4-7_vs_5-6	327	7	1:2	19	11	UCI
glass-0-4_vs_5	214	9	1:3	7	4	UCI
Kddcup	1120	41	1:99	14	9	UCI
kddcup-rootkit-imap_vs_back	2225	41	1:10	19	11	UCI
messidor_features	1200	19	1:3	420	252	KEEL
Pima	768	8	1:1	188	111	UCI
poker-8_vs_6	1477	10	1:2	13	6	UCI
Satimage	6,435	36	1:3	517	339	UCI
segment0	2,310	19	1:2	225	137	UCI
vehicle2	846	18	1:2	164	100	UCI
yeast-1-2-8-9_vs_7	947	8	1:1	20	12	UCI
yeast-2_vs_4	514	8	1:2	34	20	UCI
yeast-2_vs_8	483	8	1:2	15	9	UCI
yeast6	1,484	8	1:2	25	12	UCI
ecoli4	336	7	1:2	13	7	UCI
pendigits	10,992	16	1:3	813	473	UCI
vehicle3	846	18	1:2	151	83	UCI

- Maximum depth : None (until maximum leaf nodes are reached)
- Class weight : Balanced (to handle class imbalance)

D. Evaluation Metrics

To assess the performance of the MI-Heat algorithm and compare it with the baseline methods, several standard evaluation metrics commonly applied in imbalanced classification problems are employed:

1) PR-AUC: The area under the precision-recall curve, particularly useful for imbalanced datasets.

2) AUC-ROC: The area under the receiver operating characteristic curve, providing insight into the model's performance across different classification thresholds.

3) Partial ROC-AUC: The area under the receiver operating characteristic curve (ROC-AUC) computed within a partial range of interest (for example, considering only a certain threshold or a range of false positive rates). This metric allows for a more focused evaluation, especially in cases where specific regions of the ROC curve are of particular interest for the application.

V. RESULTS

The boxplot in Fig. 3 shows PR AUC ranks of the proposed MI-Heat on 23 benchmarking datasets with different under-sampling strategies. Lower ranks indicate better performance. As can be observed, the proposed MI-Heat method works stably well and its ranks are close to the optimal values for the most part. Traditional methods like RUS and DBSCAN are more volatile, i.e., their performances are not as stable across datasets. NearMiss and Tomek Links are superior to random under-sampling but lose some useful majority information. ENN and CNN gain better boundary cleaning at the expense of sensitivity to dataset properties. In contrast, MI-Heat strikes a good balance between throwing away redundant samples and keeping useful structure in data. As a whole, these results confirm that the new method provides more stable and accurate classification for imbalanced datasets.

Fig. 4 shows the ROC AUC rank distribution of 23 benchmark sets for different preprocessing and under-sampling methods. The ROC AUC score is used to measure the model performance in distinguishing minority and majority classes. Lower ranks show better classification performance. It can be seen from the results that the employed MI-Heat method is consistently strong in performance with relatively lower median ranks compared with traditional methods. Classic methods such as RUS and DBSCAN show higher variability, which suggests unstable performance across different datasets. CNN and ENN are suitable for boundary smoothing but show higher fluctuation, suggesting data attribute sensitivity. NearMiss and Tomek Links are fairly good but discard significant majority-class samples. Overall, the graph shows that MI-Heat produces more balanced and stable results across datasets, confirming its effectiveness in improving classifier robustness on imbalanced data.

Fig. 5 plots the distribution of Partial ROC AUC ranks across 23 benchmark datasets for different preprocessing methods. This metric focuses on the classifier's performance in the most critical region of the ROC curve, putting weight on true positive rates at low false positive rates. Lower ranks are preferable. The novel MI-Heat method has one of the lowest median ranks, reflecting stable and consistent classification performance. Methods such as RUS and DBSCAN exhibit high variability, suggesting instability across data sets. ENN and CNN present moderate performance but remain sensitive to class overlap. Near-Miss and Tomek Links improve decision boundaries but remove some informative instances. The figure 5 as a whole demonstrates that MI-Heat provides robust and stable performance within the important detection region and thus, it is highly effective for imbalanced data classification.

Fig. 6 is a ranking of Normalized Partial ROC AUC performance over 23 benchmark datasets for different preprocessing techniques. Normalized partial ROC AUC focuses on measuring performance in the most suitable part of the ROC curve with normalized results over diverse datasets of different

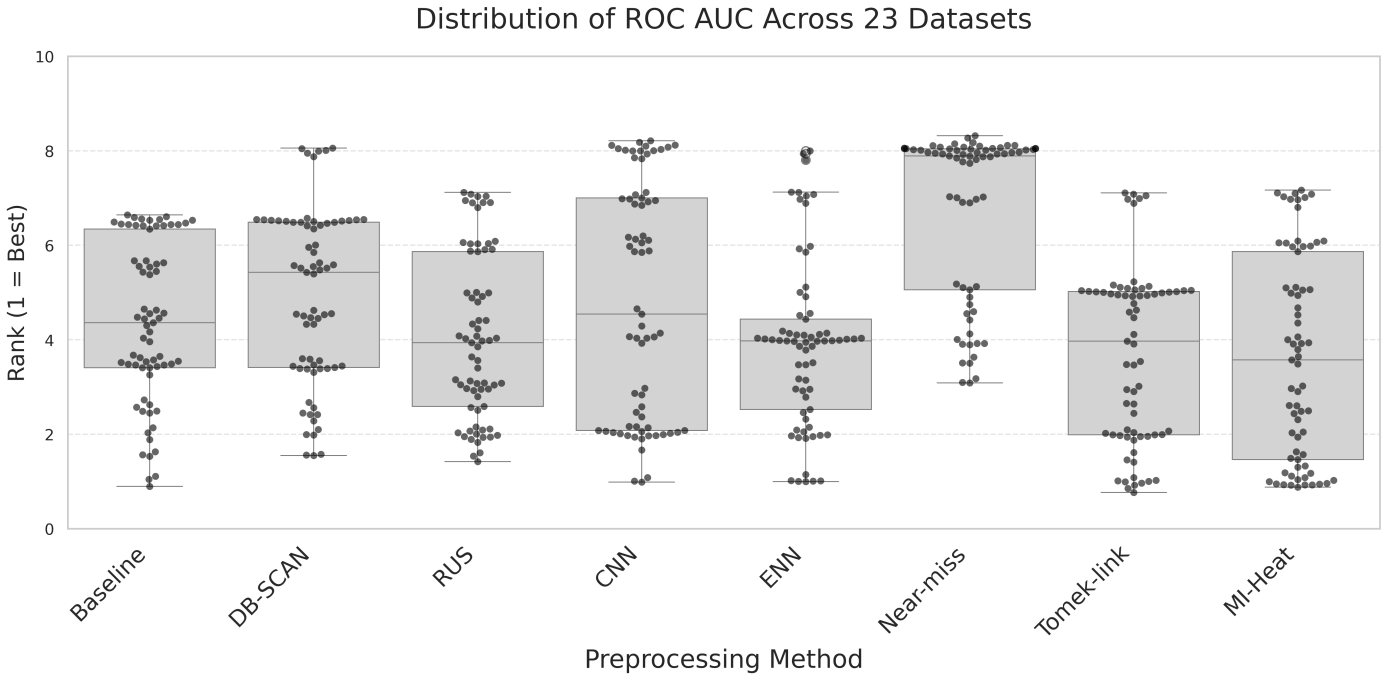


Fig. 3. PR AUC ranks of the proposed MI-Heat on 23 benchmarking datasets.

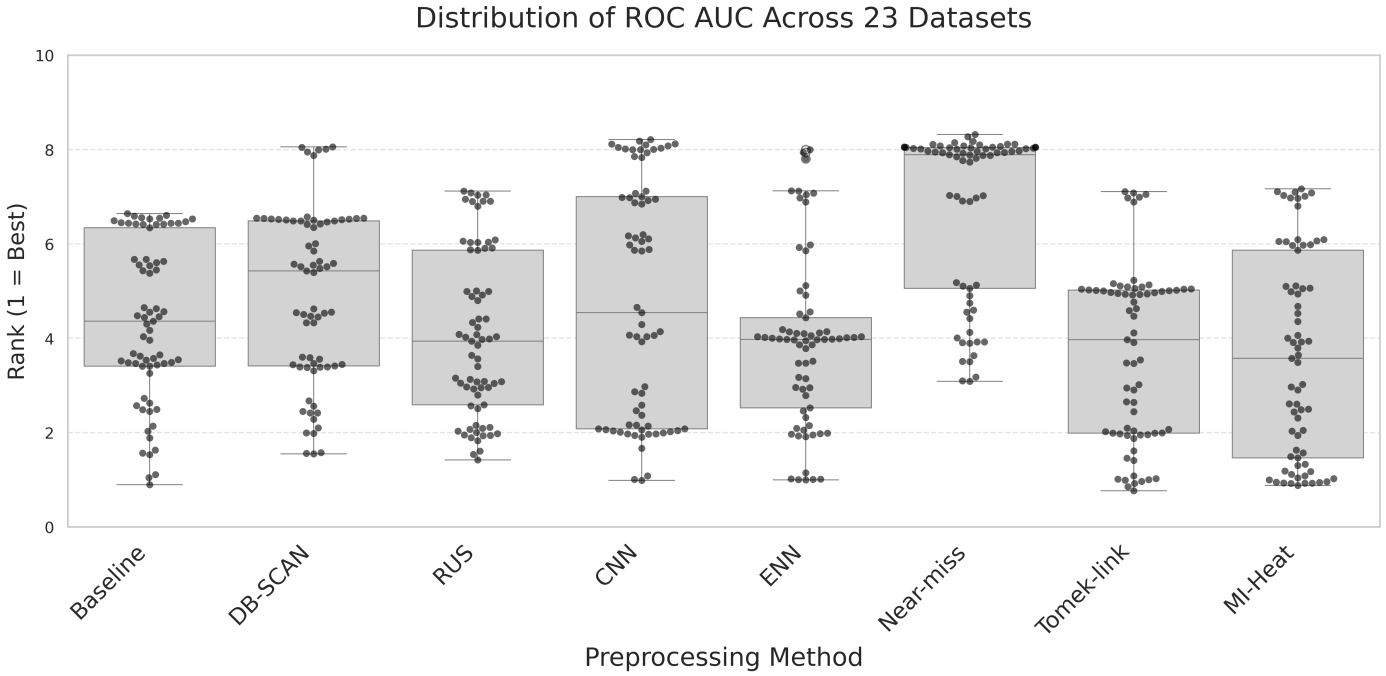


Fig. 4. ROC AUC rank distribution of 23 benchmark sets.

scales. Lower ranks correspond to superior performance. The proposed new MI-Heat method exhibits stable and competitive ranks with consistent classification quality for heterogeneous data types. Compared to traditional under-sampling approaches like RUS, DBSCAN and NearMiss, the proposed approach exhibits lower variability and higher reliability. ENN and CNN perform well at boundary refinement but are still vulnerable to overlapping classes.

When all four-measurement metrics PR AUC, ROC AUC, Partial ROC AUC and Normalized Partial ROC AUC are compared, a consistent trend can be seen. MI-Heat algorithm always yields stable, lower-ranked results, demonstrating that it is robust and resilient. It preserves data diversity well but minimizes noise to a significant degree, outperforming

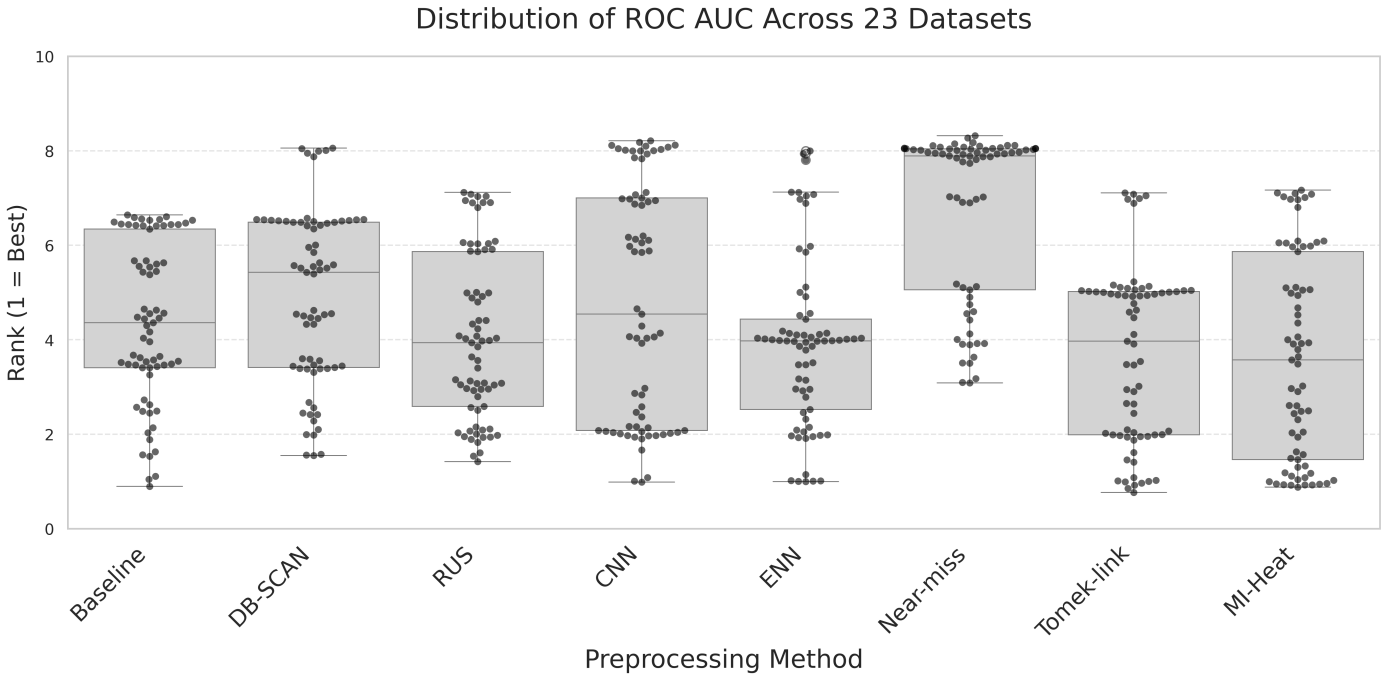


Fig. 5. Partial ROC AUC ranks across 23 benchmark datasets.

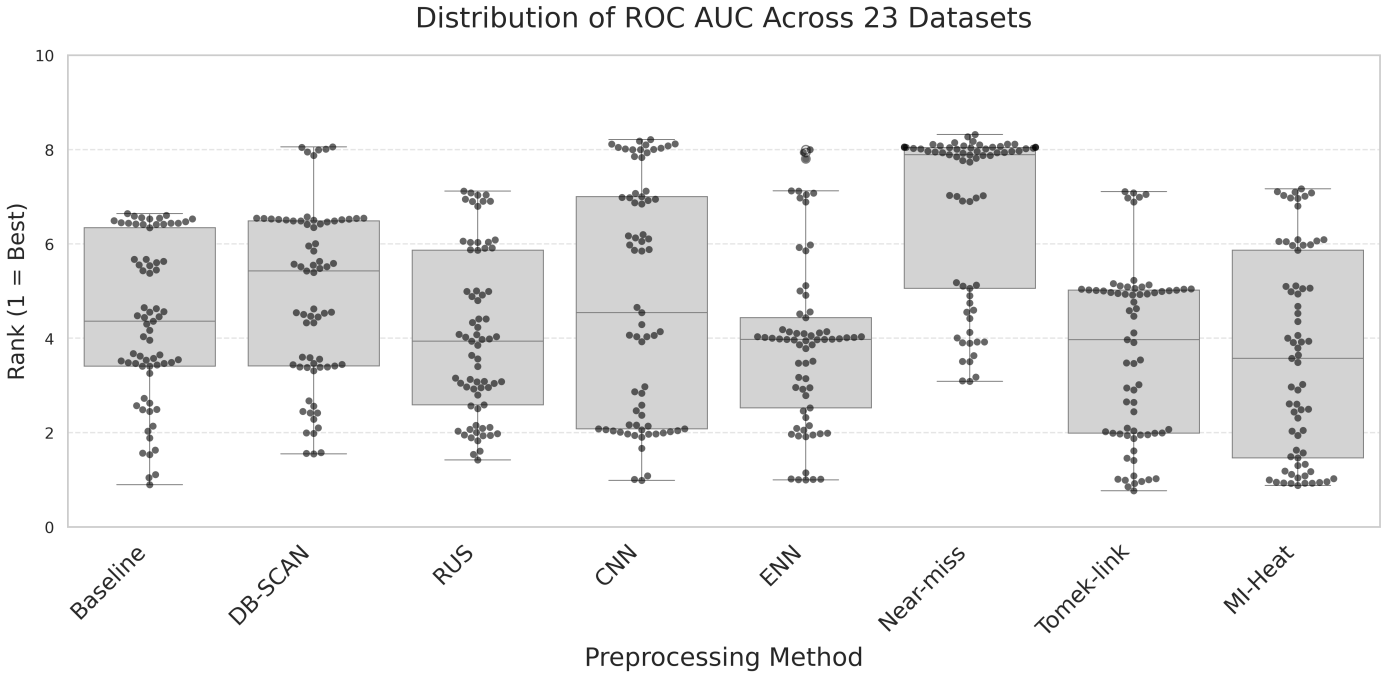


Fig. 6. Normalized partial ROC AUC performance over 23 benchmark datasets.

traditional methods which either delete valuable samples or introduce too much bias. In general, these results confirm that MI-Heat offers fair and trustworthy improvements in handling class-imbalanced datasets. The complete detailed results are provided in Tables II, IV and III and Fig. 7, and Fig. 8 (Appendix).

VI. Conclusion

This article introduced MI-Heat, a novel hybrid under-sampling algorithm incorporating Mutual Information (MI)-based feature relevance, K-Means clustering and distance heatmap visualization to effectively address class imbalance in binary classification problems. Empirical evaluation on 23 benchmark datasets demonstrated that the proposed method

achieved superior and more stable performance compared to traditional techniques such as RUS, Tomek Links, DBSCAN and NearMiss.

On all the evaluation metrics PR AUC, ROC AUC, Partial ROC AUC and Normalized Partial ROC AUC MI-Heat showed low rank variability, which signifies model robustness and stable discrimination between classes. The clustering step maintained the structural diversity of the majority class and MI-based selection ensured the retention of most informative samples. The use of the heatmap visualization added interpretability, allowing for the straightforward understanding of cluster separability and data density distribution.

Overall, the experiments confirm that MI-Heat produces an interpretable and balanced solution that minimizes data loss, avoids noise amplification and encourages classifier generalization on diverse datasets.

Meanwhile, this study has a number of limitations that need to be acknowledged. This algorithm is evaluated only on binary classification settings; multi-class or extreme-imbalance scenarios are not explored within this study. Similarly, the combination of MI computation, clustering, and heatmap generation might introduce additional computation overhead when dealing with very large datasets; comprehensive assessment of the scalability of the proposed method is beyond the scope of this study. Furthermore, given that all experiments were conducted on benchmark datasets, there is a possibility of dataset-specific biases or structural assumptions hindering the generalization in industrial settings.

VII. Future Directions

Future research can build upon this study in the following substantive ways:

- Using adaptive clustering methods like DBSCAN or Gaussian Mixture Models to automatically adapt to data density.
- Investigating multi-class imbalanced learning conditions to analyze the scalability of MI-Heat in non-binary configurations.
- Incorporating deep feature extraction or autoencoder-based representations for improving the quality of feature selection prior to under-sampling.
- Employing explainable AI (XAI) frameworks for visualizing decision boundaries and interpreting the sampling results further.
- Performing real-world domain tests in healthcare, finance and industrial fault detection to determine real-world impact.

Acknowledgment

This work was supported by the Fundamental Research Grant (Vote No. Q.K130000.3856.22H21) awarded by Universiti Teknologi Malaysia.

References

[1] Y. Jeong, K. Lee, Y. W. Park, and S. Han, "A systematic approach for learning imbalanced data: enhancing zero-inflated models through boosting," *Machine Learning*, vol. 113, no. 10, pp. 8233–8299, 2024.

[2] N. A. Mahoto, S. Gul Khan, and S. A. Ruk, "A lexicon-based method to determine subjectivity of unstructured data," in *2018 5th International Multi-Topic ICT Conference (IMTIC)*, 2018, pp. 1–7.

[3] D. Devarriya, C. Gulati, V. Mansharamani, A. Sakalle, and A. Bhardwaj, "Unbalanced breast cancer data classification using novel fitness functions in genetic programming," *Expert Systems with Applications*, vol. 140, p. 112866, 2020.

[4] R. Sewpaul, O. O. Awe, D. M. Dogbey, M. D. Sekgala, and N. Dukhi, "Classification of obesity among south african female adolescents: Comparative analysis of logistic regression and random forest algorithms," *International Journal of Environmental Research and Public Health*, vol. 21, no. 1, p. 2, 2023.

[5] P. Boulieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," *Machine Learning*, vol. 113, no. 8, pp. 5087–5108, 2024.

[6] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, and Q. Kang, "Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection," *Neurocomputing*, vol. 407, pp. 50–62, 2020.

[7] H. S. Bako, F. U. Ambursa, B. S. Galadanci, and M. Garba, "Predicting timely graduation of postgraduate students using random forests ensemble method," *FUDMA Journal of Sciences*, vol. 7, no. 3, pp. 177–185, 2023.

[8] Y. Deng and M. Li, "An adaptive and robust method for oriented oversampling with spatial information for imbalanced noisy datasets," *IEEE Access*, vol. 11, pp. 122 610–122 624, 2023.

[9] A. O. Salau, E. D. Markus, T. A. Assegie, C. O. Omeje, and J. N. Eneh, "Influence of class imbalance and resampling on classification accuracy of chronic kidney disease detection," *Mathematical Modelling of Engineering Problems*, vol. 10, no. 1, pp. 48–54, 2023.

[10] C. J. Hellín, A. A. Olmedo, A. Valledor, J. Gómez, M. López-Benítez, and A. Tayebi, "Unraveling the impact of class imbalance on deep-learning models for medical image classification," *Applied Sciences*, vol. 14, no. 8, p. 3419, 2024.

[11] J. Dunstan, F. Villena, J. P. Hoyos, V. Riquelme, M. Royer, H. Ramírez, and J. Peypouquet, "Predicting no-show appointments in a pediatric hospital in chile using machine learning," *Health Care Management Science*, vol. 26, no. 2, pp. 313–329, 2023.

[12] D. M. Tadele, G. T. Biwota, L. M. Enyew, M. M. Tadele, and G. H. Teferi, "Predicting tuberculosis incidence in adult HIV patients on ART in debre markos, ethiopia: A machine learning approach," Preprint, 2025.

[13] B. Liu and G. Tsoumakas, "Dealing with class imbalance in classifier chains via random undersampling," *Knowledge-Based Systems*, vol. 192, p. 105292, 2020.

[14] M. Loukili, F. Messaoudi, and M. El Ghazi, "Enhancing customer retention through deep learning and imbalanced data techniques," *Iraqi Journal of Science*, pp. 2853–2866, 2024.

[15] G. Chortareas, A. G. Katsafados, T. Pelagidis, and C. Prassa, "Credit risk modelling within the euro area in the COVID-19 period: Evidence from an ICAS framework," *International Journal of Finance & Economics*, vol. 30, no. 2, pp. 1074–1105, 2025.

[16] S. Ayoub, Y. Gulzar, J. Rustamov, A. Jabbari, F. A. Reegu, and S. Turaev, "Adversarial approaches to tackle imbalanced data in machine learning," *Sustainability*, vol. 15, no. 9, p. 7097, 2023.

[17] K. Huang and X. Wang, "ADA-INCVAE: Improved data generation using variational autoencoder for imbalanced classification," *Applied Intelligence*, vol. 52, no. 3, pp. 2838–2853, 2022.

[18] I. Naglik and M. Lango, "GMMSampling: a new model-based, data difficulty-driven resampling method for multi-class imbalanced data," *Machine Learning*, vol. 113, no. 8, pp. 5183–5202, 2024.

[19] F. Sağlam, "Datrel: a noise-tolerant data relocation approach for effective synthetic data generation in imbalanced classifiers," *Machine Learning*, vol. 114, no. 5, p. 126, 2025.

[20] Y. Yang, S. Huang, W. Huang, and X. Chang, "Privacy-preserving cost-sensitive learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2105–2116, 2021.

[21] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Machine Learning*, vol. 113, no. 7, pp. 4845–4901, 2024.

[22] J. Yang, R. El-Bouri, O. O'Donoghue, A. S. Lachapelle, A. A. S. Soltan, D. W. Eyre, L. Lu, and D. A. Clifton, "Deep reinforcement learning for multi-class imbalanced training: applications in healthcare," *Machine Learning*, vol. 113, no. 5, pp. 2655–2674, 2024.

[23] R. Kumar and E. Chambers, "Unreliability of clustering results in sensory studies and a strategy to address the issue," *Frontiers in Food Science and Technology*, vol. 4, 2024.

[24] T. Watthaisong, K. Sunat, and N. Muangkote, "Comparative evaluation of imbalanced data management techniques for solving classification problems on imbalanced datasets," *Statistics, Optimization & Information Computing*, vol. 12, no. 2, pp. 547–570, 2024.

[25] F. Grady, J. R. Wahidiyat, and A. S. Girsang, "Solving simulated imbalanced body performance data using A-SUWO and tomek link algorithm," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 6, no. 2, pp. 1101–1120, 2025.

[26] S. S. Kotekani and I. Velchamy, "An effective data sampling procedure for imbalanced data learning on health insurance fraud detection," *Journal of Computing and Information Technology*, vol. 28, no. 4, pp. 269–285, 2021.

[27] K. Marzuki, L. G. Rady Putra, H. Hairani, L. Z. A. Mardedi, and J. X. Guterres, "Performance improvement of the random forest method based on SMOTE-Tomek Link on lombok tourism analysis sentiment," *Jurnal Bumigora Information Technology (BITe)*, vol. 5, no. 2, pp. 151–158, 2024.

[28] W. Albattah and R. U. Khan, "Impact of imbalanced features on large datasets," *Frontiers in Big Data*, vol. 8, 2025.

[29] J. Wellnitz, S. Jain, J. Hochuli, T. Maxfield, E. Muratov, A. Tropsha, and A. Zakharov, "One size does not fit all: revising traditional paradigms for QSAR-based virtual screenings," ChemRxiv Preprint, 2023. [Online]. Available: https://doi.org/10.26434/chemrxiv-2023-9cw9v

[30] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "The effect of rebalancing techniques on the classification performance in cyberbullying datasets," *Neural Computing and Applications*, vol. 36, no. 3, pp. 1049–1065, 2024.

[31] A. Flores-Velazco and D. M. Mount, "Coresets for the nearest-neighbor rule," *arXiv preprint arXiv:2007.10440*, 2020.

[32] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[33] B. S. Arkok and A. M. Zeki, "Classification of quranic topics based on imbalanced classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 678–687, 2021.

[34] A. Salehi and M. Khedmati, "Hybrid clustering strategies for effective oversampling and undersampling in multiclass classification," *Scientific Reports*, vol. 15, no. 1, p. 3460, 2025.

[35] C. Ciciana, R. Rahmawati, and L. Qadrini, "The utilization of resampling techniques and the random forest method in data classification," *TIN: Terapan Informatika Nusantara*, vol. 4, no. 4, pp. 252–259, 2023.

[36] K. Sharma, J. Arora, P. Kherwa, and Z. Alansari, "Majority clustering for imbalanced image classification," *PeerJ Computer Science*, vol. 11, p. e2891, 2025.

[37] S. Saghafi and M. Saghafi, "Cluster-based data balancing for protein contact map prediction using MLP neural networks," Preprint, 2025.

[38] Q. Chen, T. Liu, and X. Wu, "Research on bearing zc feature selection method based on DWCMI," *Measurement Science and Technology*, vol. 35, no. 5, p. 056103, 2024.

[39] M. A. Faizin, D. T. Kurniasari, N. Elqolby, M. A. R. Putra, and T. Ahmad, "Optimizing feature selection method in intrusion detection system using thresholding," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 3, pp. 214–226, 2024.

[40] A. Gupta, C. Jain, and B. Kaur, "Evaluating machine learning models for stellar spectral classification using Gaia DR3," *Physica Scripta*, vol. 100, no. 3, p. 036016, 2025.

[41] Q. Chen, T. Liu, and X. Wu, "Research on bearing zc feature selection method based on DWCMI," *Measurement Science and Technology*, vol. 35, no. 5, p. 056103, 2024.

[42] A. H. Abu Saq, A. Zainal, B. A. S. Al-Rimy, A. Alyami, and H. A. Abosaq, "Intrusion detection in iot using gaussian fuzzy mutual information-based feature selection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17 564–17 571, 2024.

[43] S. Wang, S. Y. Kim, and K.-A. Sohn, "Clearf++: Improved supervised feature scoring using feature clustering in class-wise embedding and reconstruction," *Bioengineering*, vol. 10, no. 7, p. 824, 2023.

[44] Y. W. Tye, X. Chew, U. K. Yusof, and S. Tulpar, "ACG-SFE: Adaptive cluster-guided simple, fast, and efficient feature selection for high-dimensional microarray data in binary classification," *PLOS ONE*, vol. 20, no. 9, p. e0331089, 2025.

[45] Y. Guo, Y. Kou, L. Yi, and G. Fu, "Hibbka: A hybrid method with resampling and heuristic feature selection for class-imbalanced data in chemometrics," *Journal of Chemometrics*, vol. 39, no. 5, 2025.

[46] R. Wild, E. Sozio, R. G. Margiotta, F. Dellai, A. Acquasanta, F. D. Ben, C. Tascini, F. Curcio, and A. Laio, "Maximally informative feature selection using information imbalance: Application to COVID-19 severity prediction," Research Square Preprint, 2023.

[47] G. Baj, I. Gandin, A. Scagnetto, L. Bortolussi, C. Cappelletto, A. Di Lenarda, and G. Barbati, "Comparison of discrimination and calibration performance of ecg-based machine learning models for prediction of new-onset atrial fibrillation," *BMC Medical Research Methodology*, vol. 23, no. 1, p. 169, 2023.

[48] J. C. Munguía Mondragón, E. Rendón Lara, R. Alejo Eleuterio, E. E. Granda Gutirrez, and F. Del Razo López, "Density-based clustering to deal with highly imbalanced data in multi-class problems," *Mathematics*, vol. 11, no. 18, p. 4008, 2023.

[49] W. Zhou, C. Liu, P. Yuan, and L. Jiang, "An undersampling method approaching the ideal classification boundary for imbalance problems," *Applied Sciences*, vol. 14, no. 13, p. 5421, 2024.

[50] K. Kiruthika Devi and G. A. Sathish Kumar, "Bio-inspired algorithm based undersampling approach and ensemble learning for twitter spam detection," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 32, no. 01, pp. 1–20, 2024.

[51] A. Damayanti and A. Baita, "Comparison of support vector machine (svm) and random forest (rf) algorithm performance with random undersampling technique to predict gestational diabetes mellitus risk," *Journal of Applied Informatics and Computing*, vol. 9, no. 2, pp. 328–337, 2025.

[52] Y. Ren, D. Wu, Y. Tong, A. López-DeFede, and S. Gareau, "Issue of data imbalance on low birthweight baby outcomes prediction and associated risk factors identification: Establishment of benchmarking key machine learning models with data rebalancing strategies," *Journal of Medical Internet Research*, vol. 25, p. e44081, 2023.

APPENDIX

DETAILED SUMMARY VALUES AND RANK DISTRIBUTION

TABLE II. SENSITIVITY RESULTS OF DIFFERENT MODELS ACROSS MULTIPLE DATASETS

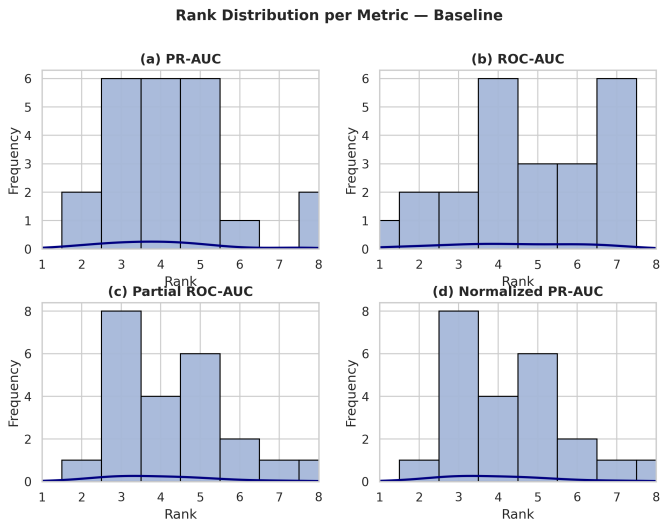
Dataset	Baseline	RUS	DBSCAN	CNN	Near Miss	ENN	Tomek Link	Hybrid
Ecoli-0_vs_1	0.955556	1	0.955556	0.955556	1	0.977778	1	0.978
Spambase	0.92721	0.961872	0.92721	0.954939	0.948007	0.965338	0.930676	0.946
Abalone20-vs-8-9-10	1	0.82716	1	0.989418	0.231041	1	1	1
Abalone9-18	0.995238	0.690476	0.995238	0.97619	0.395238	0.980952	1	1
Cardiotocography	0.981855	0.91129	0.981855	0.927711	0.796371	0.931452	0.965726	0.986
Ecoli-0-1-3-7_vs_2-6	1	0.804878	1	1	0.658537	1	1	1
Ecoli-0-1-4-7_vs_5-6	1	0.93617	1	0.93617	0.670213	1	1	1
Glass-0-4_vs_5	1	0.807692	1	1	0.961538	1	1	1
Kddcup	1	0.95842	1	0.997921	0.962578	1	1	1
Kddcup-rootkit-imap_vs_back	1	1	1	0.875758	0.918182	1	1	1
Messidor_features	0.716129	0.774194	0.716129	0.909677	0.748387	0.929032	0.798742	0.89
Pima	0.788079	0.649007	0.788079	0.615894	0.701987	0.613333	0.768212	0.907
Poker-8_vs_6	1	0.468182	1	1	0.302273	1	1	1
Satimage	0.995991	0.991404	0.995991	0.9851	0.848138	0.993123	0.994842	0.998
Segment0	1	0.996599	1	0.994898	0.945578	1	1	1
Vehicle2	1	0.985	1	0.99	1	0.98	1	1
Yeast-1-2-8-9_vs_7	1	0.785455	1	0.991727	0.276364	1	1	0.992
Yeast-2_vs_4	0.992754	0.956522	0.992754	0.985507	0.844444	0.985507	0.985507	1
Yeast-2_vs_8	0.992857	0.557143	0.992857	0.992857	0.457143	0.992857	0.992857	0.993
Yeast6	0.997706	0.885321	0.997706	0.993119	0.373853	0.995413	0.995413	1
Ecoli4	0.714286	0.893617	0.714286	1	1	1	0.666667	1
Pendigits	0.996967	0.9821	0.996967	0.9983	0.952814	0.9997	0.996967	0.997
Vehicle3	0.906736	0.6269	0.906736	0.6891	0.590674	0.7098	0.834197	0.963

TABLE III. F-MEASURE RESULTS OF DIFFERENT MODELS ACROSS MULTIPLE DATASETS

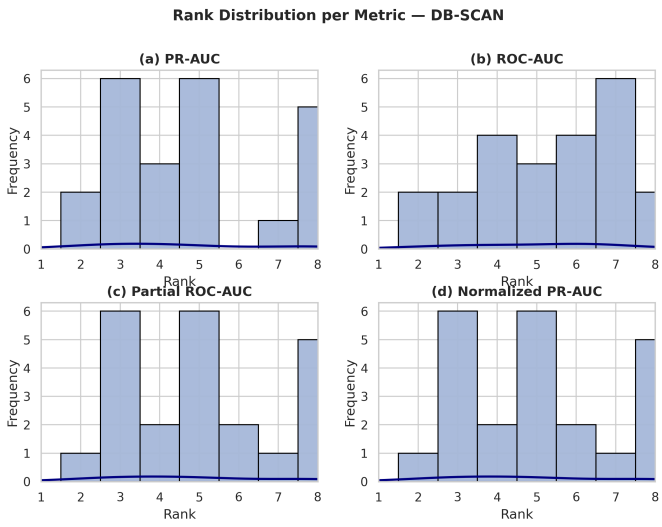
Dataset	Baseline	RUS	DBScan	CNN	Near Miss	ENN	Tomek link	Hybrid
Ecoli-0-vs-1	0.966292	1	0.966292	0.977273	0.989011	0.988764	0.988764	0.978
Spambase	0.946903	0.923461	0.946903	0.947549	0.910908	0.943268	0.938811	0.948
abalone20-vs-8-9-10	0.992126	0.903661	0.992126	0.990291	0.375358	0.992126	0.992126	0.992
abalone9-18	0.978923	0.805556	0.978923	0.973872	0.566553	0.971698	0.979021	0.979
cardiotocography	0.969154	0.944619	0.969154	0.952577	0.872928	0.953560	0.966700	0.962
ecoli-0-1-3-7_vs_2-6	0.993939	0.891892	0.993939	0.993939	0.794118	0.987952	0.993939	0.980
ecoli-0-1-4-7_vs_5-6	0.984293	0.961749	0.984293	0.961749	0.797468	0.984293	0.984293	0.978
glass-0-4_vs_5	1	0.893617	1	1	0.980392	1	1	0.981
kddcup	1	0.978769	1	0.998959	0.980932	1	1	1
kddcup-rootkit-imap_vs_back	0.999243	0.993976	0.993976	0.933764	0.957346	0.999243	0.999243	0.999
messidor_features	0.680982	0.685714	0.680982	0.689487	0.678363	0.692308	0.686486	0.685
pima	0.801347	0.734082	0.801347	0.712644	0.765343	0.710425	0.805556	0.823
poker-8_vs_6	0.995475	0.637771	0.995475	0.995475	0.463415	0.995475	0.995475	0.996
satimage	0.995991	0.994825	0.995991	0.991064	0.916692	0.995405	0.996270	0.996
segment0	1	0.998296	0.988235	0.997442	0.972028	1	1	1
vehicle2	0.997506	0.992443	0.997506	0.994975	1	0.989899	0.997506	0.995
yeast-0-2-5-6_vs_3-7-8-9	0.967509	0.908730	0.967509	0.963235	0.525199	0.967033	0.972875	0.958
yeast-2_vs_4	0.982079	0.970588	0.982079	0.985507	0.904762	0.978417	0.978417	0.974
yeast-2_vs_8	0.989324	0.715596	0.989324	0.989324	0.627451	0.989324	0.989324	0.981
yeast6	0.992018	0.935758	0.992018	0.989714	0.541528	0.990868	0.988610	0.992
ecoli4	0.989474	0.943820	0.989474	1	0.984293	0.9843	1	0.978
pendigits	0.997639	0.9908	0.997639	0.9973	0.975332	0.9917	0.997639	0.997
vehicle3	0.870647	0.758621	0.870647	0.8036	0.723810	0.7998	0.854111	0.876

TABLE IV. PR-AUC RESULTS OF DIFFERENT MODELS ACROSS MULTIPLE DATASETS

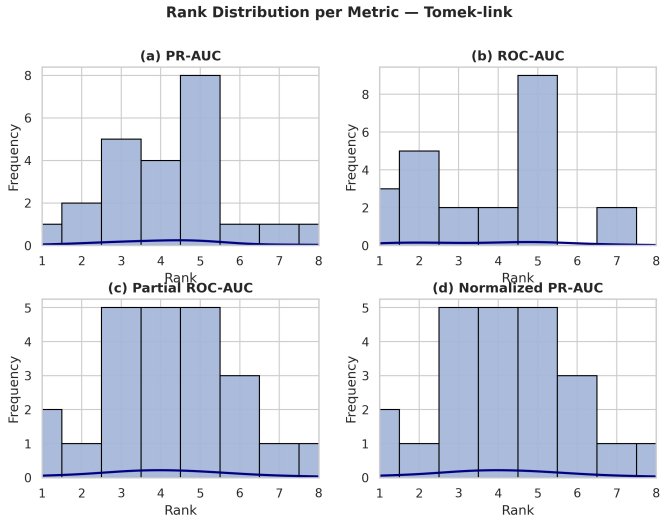
Dataset	Dbscan	Imbalanced	Near Miss	Tomek Link	Enn	Rus	Cnn	Hybrid
Poker-8_Vs_6	0.9989	0.9989	0.9946	0.9990	0.999378	0.997967	0.991607	1
Vehicle2	1	1	1	0.9998	0.999854	1	0.999926	1
Ecoli4	0.9999	0.9999	0.9986	0.9999	0.999887	1	1	1
Yeast2_Vs_4	0.9995	0.9995	0.9929	0.9993	0.973392	0.972059	0.964429	0.9996
Segmentation	0.9989	0.9990	0.9983	0.9990	1	1	0.999999	1
Abalone-9_Vs_18	1	1	1	1	0.982544	0.951903	0.984721	1
Abalone20_Vs_8-9-10	0.9793	0.9793	0.9799	0.9837	0.998056	0.998294	0.99794	0.9876
Kddcup-Land_Vs_Satan	1	1	1	1	1	1	1	1
Satimage	0.9999	0.9999	0.9985	0.9999	0.999945	0.999941	0.999705	0.9999
Pendigits	0.99998	0.99998	0.99981	0.99998	0.999973	0.999914	0.999947	0.99997
Kddcup-Rootkit-Imap_Vs_Back	0.994	1	1	1	1	1	1	1
Spambase	0.954	0.9863	0.9718	0.9851	0.983886	0.986909	0.987866	0.9883
Vehicle3	0.9528	0.9528	0.9419	0.9567	0.955635	0.952496	0.937653	0.9444
Ecoli-0_Vs_1	0.999	0.999	1	1	1	1	1	0.9995
Ecoli-0-1-4-7_Vs_5-6	0.9974	0.9974	0.9662	0.9974	0.998175	0.998633	0.994141	0.9978
Ecoli-0-1-3-7_Vs_2-6	0.9999	0.9999	0.9994	0.9999	0.999703	0.999627	1	0.9992
Yeast-0-2-5-6_Vs_3-7-8-9	0.9995	0.9995	0.9928	0.9993	0.973392	0.972059	0.964429	0.9996
Cardiography	0.9968	0.9968	0.9975	0.9960	0.994158	0.994949	0.994911	0.9967
Messidor_Features	0.7019	0.7019	0.7111	0.6897	0.681141	0.712438	0.656291	0.6706
Pima	0.8841	0.8841	0.8791	0.8862	0.874282	0.873859	0.877073	0.887
Yeast6	0.9955	0.9955	0.9742	0.9957	0.99669	0.997035	0.993282	0.9945
Yeast-2_Vs_8	0.9979	0.9979	0.9975	0.9960	0.997831	0.992736	0.997611	0.9969
Glass-0-4_Vs_5	0.9998	1	1	1	1	1	1	1



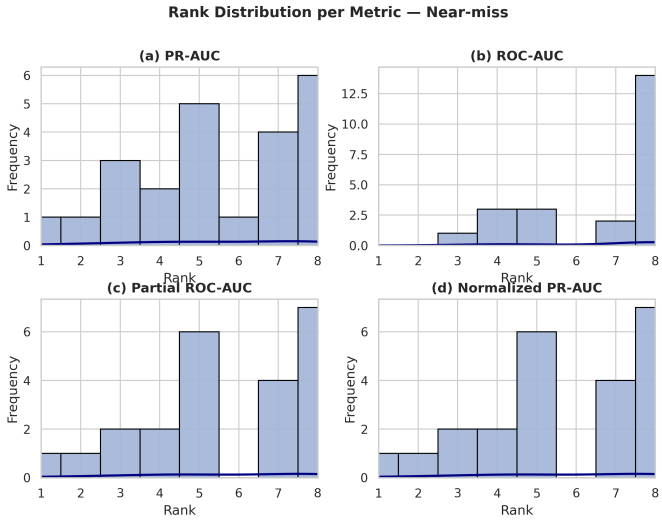
(a) Baseline



(b) DB-SCAN

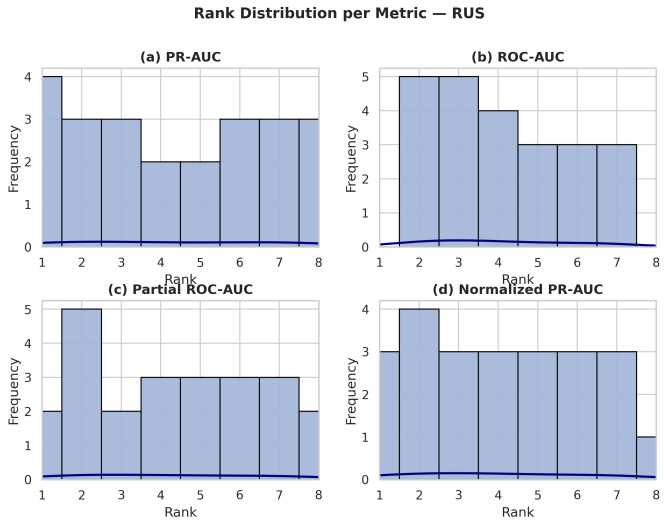


(c) Near-Miss

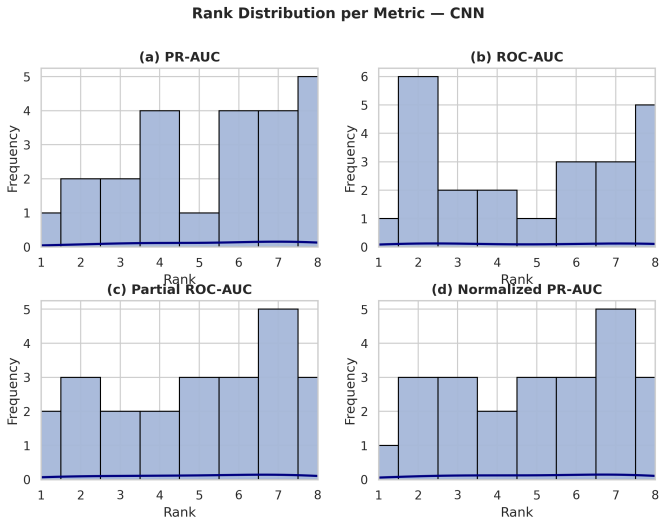


(d) Tomek-link

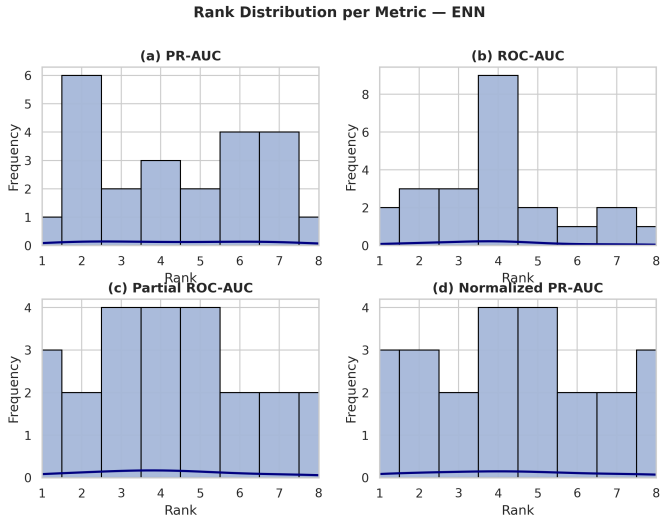
Fig. 7. Rank distribution for the first four preprocessing methods.



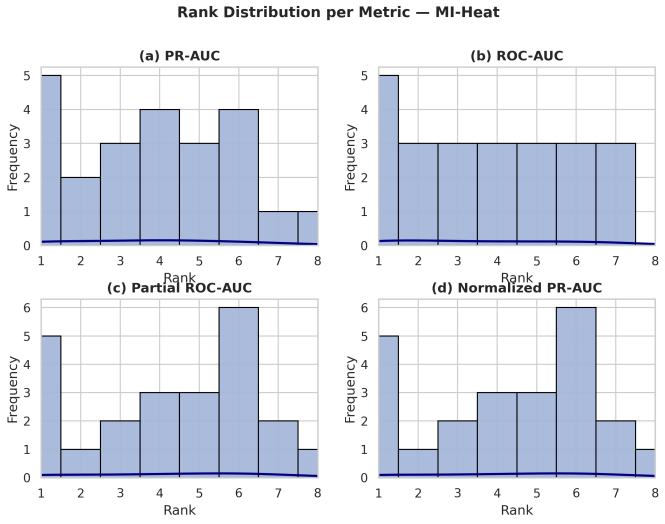
(a) CNN



(b) ENN



(c) RUS



(d) MI-Heat

Fig. 8. Rank distribution for the remaining preprocessing methods.