Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084436

ARTICLE

A Novel Entropy-Based Framework for Hybrid Sampling in Imbalanced Learning

Ren-Jieh Kuo[1,*], Muhammad Rizki[1], Ferani Eva Zulvia[2] and Eddy Roflin[3]

[1]Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan
[2]Department of Industrial Engineering, Institut Teknologi Bandung, Bandung, Indonesia
[3]Faculty of Medicine, Sriwijaya University, Palembang, Indonesia

*Corresponding Author: Ren-Jieh Kuo. Email: rjkuo@mail.ntust.edu.tw

Received: 22 April 2026; Accepted: 01 July 2026; Published: 13 August 2026

ABSTRACT: Imbalanced data remain a critical challenge in classification, as skewed distributions bias models toward majority classes and diminish sensitivity to minority classes, which are often the most critical. To address this issue, this paper proposes the Information Filtered Hybrid Algorithm (IF-HA), a novel entropy-based sampling method that integrates undersampling and oversampling guided by information theory. IF-HA quantifies instance importance through an instance-wise difference statistic. In the undersampling stage, majority of instances with low difference statistics in the border area are eliminated, while in the oversampling stage, synthetic samples are generated from two minority core points or two minority instances with high difference statistics located in the border area. This process removes noise, eliminates redundant majority border points, and generates synthetic minority samples in informative regions until an entropy-based imbalance threshold is reached. The proposed algorithm is evaluated on 20 benchmark datasets from the UCI and KEEL repositories. Results demonstrate that IF-HA consistently improves minority detection and achieves higher F1 Scores, recall, and AUC (Area Under the Curve) than other methods, including SMOTE, Borderline-SMOTE, ADASYN (Adaptive Synthetic Sampling), and SMOTE-TLNN-DEPSO. A real-world tuberculosis (TB) dataset from Indonesia was further used to validate the practical applicability of IF-HA using KNN, Random Forest, and XGBoost (eXtreme Gradient Boosting) classifiers. The results show consistent improvements after applying IF-HA. These findings indicate that entropy-based hybrid sampling is a promising approach for structured tabular imbalanced classification, while further validation on high-dimensional text and image datasets remains necessary to establish broader generalizability.

KEYWORDS: Imbalanced data; hybrid sampling; information theory; entropy-based algorithm; minority class detection; tuberculosis classification

1 Introduction

Imbalanced data remains a major challenge in classification tasks. When one class significantly outnumbers others, learning algorithms tend to be biased toward the majority class, leading to high overall accuracy but poor recognition of minority instances. This bias undermines model reliability in domains where the minority class is often the most critical, such as fraud detection or rare disease diagnosis. However, numerous methods have been proposed to address this issue, ranging from resampling techniques to algorithmic modifications [1–3]. Imbalanced classification continues to demand new strategies that improve minority detection without sacrificing general performance.

Algorithms for handling imbalanced data are generally divided into four categories: data preprocessing, cost sensitive learning, algorithm level modifications, and hybrid approaches. The preprocessing methods rebalance the dataset before training. The cost sensitive learning assigns higher penalties to minority misclassifications. The algorithm level modifications adapt model mechanics. The hybrid approaches that combine these strategies to exploit their complementary strengths and improve classification performance on imbalanced dataset [4]. Preprocessing is the most direct approach to address class imbalance, as it modifies the training data before model construction. The goal is to balance class distributions so that classifiers treat all classes more equitably. Two common strategies are oversampling and undersampling. The oversampling increases the number of minority samples, while undersampling reduces the majority samples by selectively removing redundant or noisy instances. These techniques can also be combined to mitigate their individual drawbacks. Recent advances extend preprocessing with deep oversampling, which integrates synthetic sample generation with deep neural networks to capture complex feature patterns more effectively [5]. By enriching minority representations while retaining majority diversity, preprocessing methods provide a flexible, model agnostic foundation for improving fairness and predictive accuracy in imbalanced classification.

While the preprocessing approach modifies the data to balance class distribution, cost sensitive and algorithm centered approaches operate at the algorithm level. These strategies adjust the algorithm's behavior to address data imbalances by changing how classification errors are penalized or by directly altering the algorithm's mechanics to recognize the minority class better. Cost sensitive learning assigns higher penalties to misclassifications of the minority class, thereby adjusting the learning process to reduce such costly errors. This concept can be integrated into many existing classification algorithms, enhancing their effectiveness in dealing with imbalanced data. Integrating cost sensitive learning into classification algorithms allows them to account for the consequences of misclassifying different classes. This is especially valuable in scenarios with imbalanced datasets, where the risk of misclassifying minority-class data is typically higher. By assigning more significant penalties to these errors, the algorithm can adjust its focus and improve accuracy in recognizing and classifying instances from underrepresented classes [6–9]. On the other hand, the algorithm centered approach involves more profound modifications compared to the cost sensitive approach. It tailors the internal workings of classification algorithms to specifically improve how they handle imbalanced data, focusing on better recognition and processing of the minority class. For instance, bagging or ensemble algorithms enhance classification accuracy by combining multiple models, which helps in averaging biases and improving the overall performance across diverse data scenarios. Indeed, by leveraging the strengths of multiple algorithms through techniques like bagging or ensemble methods, the approach creates a more robust and accurate model. This is particularly effective for managing the complexities of imbalanced datasets, where such strategies can significantly improve how the system handles data diversity, reducing bias towards the majority class while enhancing sensitivity to the minority class [10,11]. Each approach to handling imbalanced data classification has its advantages and drawbacks. While cost sensitive learning and algorithm centered modifications can significantly improve the prediction accuracy for the minority class, they may introduce added complexity and require more computational resources. On the other hand, methods like bagging and ensemble techniques enhance the robustness of models but often require more data and lengthier training periods, which can be resource intensive. These trade offs must be considered when selecting the most appropriate strategy for a given dataset. Indeed, it is crucial to balance these factors when selecting the most suitable approach for dealing with imbalanced data, depending on the specific application. Considerations such as the extent of data imbalance, the computational resources available, and the criticality of accurately predicting minority class instances guide the decision on whether to use cost

sensitive learning, algorithm modifications, or ensemble methods. This strategic decision making ensures that the chosen method aligns well with the overall goals and constraints of the project [12].

Despite years of research, imbalanced data classification continues to present unresolved challenges and opportunities for innovation. These persistent issues fuel ongoing research efforts, as existing methods still offer significant potential to address the complexities of imbalanced datasets. This dynamic field encourages the development of new approaches and the refinement of existing techniques to enhance model accuracy and applicability across diverse scenarios [13]. Imbalanced datasets are especially common in areas such as finance, where events like fraud occur much less frequently than regular transactions. This disparity makes it difficult to accurately predict these rare events. Imbalanced datasets also frequently arise in physical, industrial, and medical systems, where collecting large numbers of instances is costly, time consuming, or practically infeasible. Examples include rare fault detection in manufacturing lines, sensor-based monitoring of critical equipment, and medical imaging of uncommon conditions. In these scenarios, traditional resampling or algorithmic approaches may fail to effectively capture minority class patterns. Therefore, developing more effective methods for managing such data imbalances is critical to improving prediction accuracy and ensuring reliable outcomes in real world applications [14], medical [15–17], networks [18–20], agriculture [21,22], and others [23–25].

This study aims to propose a novel hybrid-sampling algorithm, named the information filtered hybrid algorithm (IF-HA). The proposed algorithm addresses the imbalance by integrating undersampling and oversampling under the guidance of information theory. The main idea of the proposed algorithm is to eliminate the majority of unimportant instances and generate synthetic minority instances based on significantly important minority instances. The IF-HA algorithm utilizes an instance-wise difference statistic in information theory to quantify the importance of each data point. In the first stage, the algorithm removes the noise to reduce distortion. Then, it iteratively eliminates low-value border points from the majority class and synthesizes minority samples around informative core and border regions. This process continues until an entropy-based imbalance degree reaches a balanced threshold, ensuring that the resulting dataset is both representative and discriminative. By combining principled filtering with targeted sample generation, IF-HA enhances robustness and accuracy while preserving the utility of the original data.

The rest of this study is organized as follows. Section 2 provides the literature survey for the related topics. Section 3 presents the proposed hybrid algorithm, while the experimental results are shown in Section 4. Finally, the concluding remarks are made in Section 5.

2 Literature Review

The proposed IF-HA algorithm is grounded in fundamental concepts of imbalanced data classification, including oversampling, undersampling, hybrid strategies, and information theory. This section provides a brief review of the underlying theoretical foundations relevant to the present study.

2.1 Imbalanced Data Classification

One common challenge in classification tasks is dealing with imbalanced datasets, where one or more dominant classes, known as majority classes, significantly outnumber the minority classes [26]. This imbalance can lead to biased classification models that fail to recognize patterns in the minority class adequately [27]. Addressing this issue requires the classification algorithm to account for both majority and minority classes, ensuring sufficient representation and learning from the minority data. Solutions to this problem include transforming the imbalanced dataset into a balanced one through preprocessing methods or altering the classification algorithm. This study focuses on preprocessing techniques for handling imbalanced

data. While hybrid preprocessing techniques can mitigate imbalance at the data level, evaluating model performance remains a critical challenge.

2.2 Oversampling Approaches

Oversampling techniques have proven to be effective data augmentation methods for addressing class imbalance in machine learning. By generating additional data points for the minority class, these approaches help balance class distribution and improve classification performance. Among these techniques, methods that generate synthetic data have gained significant attention, particularly the Synthetic Minority Oversampling Technique (SMOTE) introduced by [28]. Unlike simple duplication, SMOTE creates synthetic samples by interpolating between minority class instances and their k nearest neighbors. This mechanism expands the coverage of the minority class, reduces overfitting, and improves model generalization. As a result, SMOTE has become a fundamental baseline method for handling imbalanced datasets.

To enhance SMOTE, Han et al. [29] proposed Borderline-SMOTE, which focuses on minority instances near the decision boundary, known as borderline instances. These instances are critical because they are more likely to be misclassified. Borderline-SMOTE improves performance by generating synthetic samples primarily from the danger region, where minority samples are surrounded by majority class neighbors. To achieve this, minority samples are categorized into three groups: safe samples (mostly surrounded by minority neighbors), danger samples (surrounded by a mix of classes), and noise samples (surrounded entirely by majority class neighbors). This targeted sampling strategy enables better refinement of decision boundaries and improves classification accuracy compared to SMOTE and random oversampling.

Building upon these ideas, several extensions have been developed to further improve oversampling performance. For example, Affinitive Borderline SMOTE (AB-SMOTE) incorporates affinity measures to refine the selection of borderline samples [30], while Hybrid Clustered Affinitive Borderline SMOTE (HCAB-SMOTE) combines oversampling and undersampling to improve efficiency and classification results [31]. Advanced SMOTE (A-SMOTE) optimizes the placement of synthetic samples based on proximity relationships [32]. In addition, the Adaptive Synthetic (ADASYN) method generates synthetic samples adaptively by assigning higher importance to minority instances that are more difficult to classify [33]. Collectively, these methods demonstrate the evolution of oversampling techniques in addressing imbalanced data challenges. They have been successfully applied in various domains, such as medical diagnosis [34], where minority cases represent rare diseases, and fraud detection [30], where fraudulent transactions are significantly outnumbered by normal ones [35].

2.3 Undersampling Approaches

Undersampling is a technique used to address class imbalance by reducing the size of the majority class through the removal of instances, thereby creating a more balanced dataset. Among these approaches, K-nearest neighbor (KNN)-based undersampling methods focus on eliminating redundant, noisy, or borderline majority class instances by analyzing their relationships with neighboring samples. One-sided selection, proposed by Kubat and Matwin, combines Tomek Links and the Condensed Nearest Neighbor (CNN) rule, where Tomek Links identify ambiguous instances near class boundaries for removal, and CNN eliminates redundant majority samples far from the boundary. This concept was further extended by Hart [36] and Laurikkala [37] through the Neighborhood Cleaning Rule, which systematically removes noisy or borderline instances. More recent developments include enhanced methods integrating nearest neighbor and Tomek Links techniques [38], as well as ranking-based approaches that prioritize majority instances for removal using weighted Euclidean distance to minority samples [39]. Despite their effectiveness, these methods face

limitations, particularly the lack of control over the number of removed instances, which depends heavily on dataset characteristics and may result in insufficient balancing.

Another widely used approach is K-means-based undersampling, which leverages clustering techniques to represent the majority class more effectively. Yen and Lee [40] introduced a method that clusters majority class instances and randomly selects representative samples from each cluster to maintain diversity. Chen and Shyu [41] extended this idea by forming K-balanced subsets, pairing clustered majority groups with minority data to train multiple subspace models, thereby improving robustness. Lin et al. [42] further refined this method by increasing clustering granularity to ensure more representative data selection. In many implementations, cluster centroids are used to replace original majority instances, creating a compact and balanced training set. Additionally, Shobana and Prakash Battula [43] proposed an advanced clustering-based undersampling method that removes rare instances and outliers using diversified distribution strategies, improving data quality and reducing noise.

However, K-means-based undersampling methods also exhibit several limitations. They assume spherical cluster structures, which may not accurately represent complex data distributions, and require careful selection of the parameter K, which significantly influences performance. Furthermore, these methods often fail to adequately address class overlap, potentially reducing classification effectiveness. To overcome these challenges, alternative undersampling techniques have been proposed. For example, the radial-based undersampling method [44] employs a non-nearest neighbor strategy to better handle outliers and small disjunctions. Similarly, the progressive density-based undersampling approach by Xie et al. [45] iteratively extracts majority instances based on density peaks to improve data quality. Another approach, proposed by Ha and Lee [46], utilizes evolutionary algorithms to optimize instance selection. Collectively, these methods provide more flexible and effective solutions for handling complex imbalanced datasets.

2.4 Hybrid Approaches

The hybrid approach is a data-level technique designed to address class imbalance by combining the strengths of both over-sampling and under-sampling methods. This strategy aims to balance datasets by simultaneously reducing majority class samples and increasing minority class instances, thereby overcoming the limitations of using either method independently [47]. To mitigate these drawbacks, hybrid techniques integrate the benefits of both approaches, enabling more effective data processing by selectively decreasing majority samples while generating informative minority samples. For example, Sáez et al. [48] addressed the noise issue in SMOTE by combining it with under-sampling techniques such as Edited Nearest Neighbors (ENN) and Tomek Links to remove noisy or overlapping samples. In another study, reference [49] proposed a spatio-temporal hybrid algorithm that integrates over-sampling and selective down-sampling to handle imbalance in video image processing, particularly for foreground–background segmentation. Additionally, reference [50] introduced a dual particle swarm optimization method to simultaneously perform under-sampling and over-sampling, achieving efficient dataset balancing within a shorter computational time. These hybrid strategies demonstrate flexibility and robustness in handling class imbalance across various applications.

Various hybrid sampling approaches further incorporate specialized mechanisms to detect and eliminate noise in imbalanced datasets. Techniques such as SMOTE-ENN and SMOTE-Tomek Links (SMOTE-TL), introduced by [51], leverage local data characteristics to identify and remove noisy or ambiguous instances, where ENN eliminates samples with inconsistent nearest neighbors and Tomek Links removes borderline instances between classes. SMOTE-RSB [52] integrates rough set theory by applying lower approximation concepts to detect and eliminate noisy samples, ensuring a more accurate representation of the minority class. Meanwhile, SMOTE-IPF, proposed in [48], employs an iterative under-sampling

framework to progressively identify and discard noisy data points, continuously refining the dataset. Collectively, these advanced hybrid methods significantly improve the handling of imbalanced datasets by reducing noise while preserving meaningful data distributions, ultimately enhancing classification performance.

2.5 Information Theory

Information theory, a fundamental concept in data science and machine learning, is crucial in addressing challenges posed by imbalanced data. Imbalanced data occurs when the distribution of classes in a dataset is significantly skewed, often leading to biased models that favor the majority class. C. E. Shannon introduced information entropy, a measure of uncertainty or randomness in a system [53]. In the context of imbalanced data, entropy quantifies the degree of class distribution imbalance, reflecting lower diversity or variability in class representation. This insight can help guide the selection and adjustment of machine learning models to handle such imbalances better. Additionally, information gain, derived from entropy, measures the reduction in uncertainty after splitting a dataset based on an attribute. It is beneficial in assessing the discriminative power of features in imbalanced datasets, aiding in feature selection and enhancing model performance.

Recent studies have utilized information theory to tackle imbalanced data problems. Examples include the entropy-based matrix learning machine [54], the entropy-based imbalance degree (EID) [55], and the entropy-based time-series model [56]. These approaches leverage entropy to characterize the uncertainty and structure of systems, providing a foundation for effective model development. As mentioned by Li et al. [57], Entropy is a versatile metric for describing information content and is applicable across diverse fields and use cases. This adaptability underscores its value in addressing the complex challenges associated with imbalanced data.

3 Research Method

The classification of imbalanced data is a significant research area due to its widespread occurrence in practical applications such as healthcare and manufacturing. As previously mentioned, data-level algorithms for handling imbalanced datasets can be categorized into three primary approaches: over-sampling, under-sampling, and hybrid-sampling. This study focuses on hybrid sampling and aims to develop a novel algorithm to balance imbalanced datasets, thereby enhancing the effectiveness of existing classification methods. The proposed method employs information theory to evaluate the relative importance of individual data points. Using this information, the algorithm can generate synthetic minority data while reducing majority class data. To implement and evaluate the proposed approach, a structured research workflow is designed as follows.

The research process begins with the collection of imbalanced data, which includes both benchmark datasets and real-world problem datasets. The collected data then undergo preprocessing using a hybrid technique that combines oversampling and undersampling to achieve a more balanced class distribution. Based on the newly balanced data, a classification model is developed using the K-Nearest Neighbors (KNN) algorithm. The performance of the model is subsequently evaluated and validated using the available datasets to ensure its effectiveness. Finally, the proposed approach is implemented on a real-world dataset, specifically in the medical domain, to demonstrate its practical applicability, after which the process is concluded.

3.1 Data Collection

Data collection represents the initial phase, during which raw data is gathered from various sources for subsequent data analysis and model development stages. The proposed algorithm will undergo validation using 20 benchmark datasets, which will be sourced from publicly available databases like the KEEL and UCI repositories. The proposed algorithm is also applied to a real-world dataset, collected from patient data

of tuberculosis (TB) cases in Indonesia. The tuberculosis dataset is primary data obtained through a survey involving respondents and was provided by the Faculty of Medicine, Sriwijaya University. The study protocol, informed consent procedure, and data collection process were approved by the Ethics Committee of the Faculty of Medicine, Sriwijaya University (Approval No. 0068/UN9/SB.SU/2020). All participants provided written informed consent. The dataset was collected specifically for this study and is not publicly available due to ethical and confidentiality considerations. The increasing prevalence of TB, particularly during the COVID-19 pandemic, has exacerbated public health challenges and imposed significant economic burdens on communities. Healthcare institutions have been striving to address this situation by providing diagnostic, treatment, and preventive services to patients across various demographics. These services are critical for managing TB cases effectively and reducing its transmission within communities.

A major challenge healthcare institutions face is ensuring patients adhere to their prescribed treatment regimens. Non-adherence to TB treatment, often due to negligence, lack of awareness, or socio-economic barriers, leads to higher rates of drug resistance and prolonged infectiousness. This strains healthcare systems and undermines efforts to control the disease. Therefore, identifying patients likely to adhere to their treatment regimens is essential. This can be achieved through classification methods.

The TB dataset contains 480 instances with an IR 2.43. It has twelve features: weight, height, body mass index, age, sex, education level, nutritional status, economic status, smoking habits, household contacts, family size, and BCG (Bacillus Calmette-Guérin) immunization. The output is divided into two categories, with 140 and 340 instances for each category, respectively.

3.2 Data Preprocessing

Data pre-processing is a critical step in classification, as raw data often contains inconsistencies, errors, irrelevant information, and imbalances that can negatively impact model performance. This phase typically involves data cleaning, handling missing values, encoding categorical variables, and normalizing or scaling numerical features.

For imbalanced datasets, pre-processing may include balancing techniques such as removing instances from the majority class or generating additional samples for the minority class. Majority class handling can be achieved using under-sampling techniques, such as Random UnderSampling (RUS), which reduces the number of instances in the majority class to achieve balance. However, under-sampling must be performed carefully to avoid losing valuable information.

To address the imbalance of minority data, oversampling methods like the Synthetic Minority Over-sampling Technique (SMOTE) are often employed. SMOTE generates synthetic samples by interpolating between the closest neighbors of the minority class, effectively increasing its representation while preserving the dataset's integrity.

3.3 The Proposed Algorithm

The entropy in information theory indicates the importance of the "information" contained in the data. A dataset with higher entropy carries more "information" than a dataset with lower entropy. In the imbalanced dataset, entropy can be applied to balance the information between the majority and minority classes. This study employs an entropy based density formulation inspired by Li et al. [58] and adapts it to evaluate instance contribution in the proposed hybrid sampling framework. Given a dataset $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$ in $\boldsymbol{D}$-dimensional space with $\boldsymbol{M}$ classes. Let $\boldsymbol{N_c}$ is the number of data points belonging to the class $\boldsymbol{C}$. For each data point $\boldsymbol{x_i}$, the density-based instance-wise statistic of $\boldsymbol{x_i}$, $\boldsymbol{\lambda}(\boldsymbol{x_i})$, can be defined as in Eq. (1):

$$\lambda(x_i) = \begin{cases} \dfrac{1}{|K_i|} \displaystyle\sum_{l=1, x_l \varepsilon K_i}^{|K_i|} \dfrac{1}{d(x_i, x_l)}, & if\ K_i \neq \varnothing 0, \quad if\ K_i = \varnothing, \end{cases} \tag{1}$$

where K_i is a set of the k-nearest data points, and $|K_i|$ is the size of set K_i.

The entropy-based statistical measure for class c, represented as θ_c, is determined using Eq. (2) as follows:

$$\theta_c = -\frac{1}{N_c} \sum_{i=1}^{N_c} \gamma_i \gamma_i, \tag{2}$$

where γ_i is a percentage of the overall density metric for x_i in class c and defined in Eq. (3):

$$\gamma_i = \frac{\lambda(x_i)}{\sum_{j=1, j \in C}^{N_c} \lambda(x_j)}. \tag{3}$$

The required additional information content of data x_i in class c, notated as $\delta\left(\theta_c || \vartheta_c^i\right)$, is calculated using Eqs. (4) and (5) as follows:

$$\delta\left(\theta_c || \vartheta_c^i\right) = \frac{-\sum_{j=1}^{N_c} \gamma_j \gamma_j}{N_c} log_2 \frac{-\frac{1}{N_c} \sum_{j=1}^{N_c} \gamma_j \gamma_j}{-\frac{1}{N_c - |K_i|} \sum_{j=1, j \notin K_i}^{N_c} \gamma_j^i \gamma_j^i} and \tag{4}$$

$$\gamma_j^i = \frac{\lambda(x_j)}{\sum_{j=1, j \notin K_i}^{N_c} \lambda(x_j)}. \tag{5}$$

The instance-wise difference statistic of x_i in class c for the entire dataset, denoted by $\varpi\left(\theta_c || \vartheta_c^i\right)$, is calculated using Eq. (6). The proposed instance-wise difference statistic is conceptually related to information theoretic divergence measures. Eq. (4) quantifies the entropy variation produced by removing the local neighborhood contribution of a data point. This mechanism resembles localized distributional divergence estimation, where larger values indicate that the instance contributes more strongly to the structural information of the class distribution.

$$\varpi\left(\theta_c || \vartheta_c^i\right) = \frac{e^{\delta\left(\theta_c || \vartheta_c^i\right)}}{\sum_{t=1}^{N} e^{\delta\left(\theta_c || \vartheta_c^t\right)}}. \tag{6}$$

Eq. (6) provides the standalone formal definition of the entropy-based instance-wise importance score used throughout IF-HA. Let $\varpi\left(\theta_c || \vartheta_c^i\right)$ denote the importance score assigned to instance x_i belonging to class c. The score is derived from the additional information content $\delta\left(\theta_c || \vartheta_c^i\right)$, which quantifies the entropy variation caused by excluding the local neighborhood contribution of x_i. Higher values indicate that the instance contributes more strongly to the information structure of its class, whereas lower values indicate redundant or less informative instances. Consequently, majority class border instances with low scores are prioritized for undersampling, while minority class border instances with high scores are prioritized for synthetic sample generation.

Although the proposed formulation is not equivalent to classical mutual information, it follows the same principle of measuring information change between two probability structures. In this study, the entropy variation is utilized as a practical importance indicator for identifying informative border and core instances during iterative sampling.

For each class c, the class-wise difference statistic, η_c, is defined in Eq. (7) as follows:

$$\eta_c = \frac{1}{N_c}\sum_{i=1}^{N_c} \omega\left(\theta_c \| \vartheta_c^i\right). \tag{7}$$

The entropy-based imbalanced degree for the dataset X, $EID(X)$, is given by Eq. (8):

$$EID(X) = \frac{1}{M}\sum_{c=1}^{M} |\eta_c - \xi|, \tag{8}$$

where

$$\xi = \frac{1}{M}\sum_{c=1}^{M} \eta_c. \tag{9}$$

This study applies the entropy formula defined in Eqs. (1)–(9) to identify high-important and low-important data points. It is also used as an indicator of an imbalanced ratio. The proposed algorithm utilizes information theory to improve the imbalanced ratio in the dataset.

The main ideas of the proposed entropy-based hybrid algorithm (IF-HA) are as follows:

(1) Determine the importance level of each data point using an entropy,
(2) Identify core, border, and noise points for majority and minority classes,
(3) Eliminate noise points,
(4) Iteratively eliminate border data from the majority class and generate artificial data for the minority class until the imbalanced ratio is acceptable.

The proposed IF-HA algorithm (Algorithm 1) is presented as follows and illustrated in Fig. 1.

Algorithm 1: Entropy-based hybrid algorithm

INPUT: $X = \{x_1, \ldots, x_N\}$ in D-dimensional space with M classes. Parameter ε, which controls the final imbalanced quality and β, which controls the oversampling procedure.

STAGE 1: Initial filtering

1) Identify each class's core, border, and noise data points.
2) Eliminate noise points in all classes.

STAGE 2: Under-sampling and over-sampling iteratively

Calculate the Entropy-Based Imbalance Degree (EID) using Eq. (8).
Calculate instance-wise difference statistics of all remaining data points using Eq. (6).

WHILE $EID \geq \varepsilon$ **DO**

Under-sampling process
Delete a border data point belonging to the majority class with the minimum instance-wise difference statistic.
Over-sampling process
Generate a random number $r \sim U(0,1)$.
IF $r \leq \beta$ **THEN**
Select two random core points from the minority class, x_i and x_j.
Generate artificial data points $x_k = x_i + rand(0,1)\left(x_i - x_j\right)$.
ELSE
Find two border points from the minority class, x_i and x_j which have the highest instance-wise difference statistic values.

(Continued)

Algorithm 1 (continued)

Generate artificial data points: $x_k = x_i + 0.5r\ and\ (0,1)\left(x_i - x_j\right)$.

END IF

Calculate the Entropy-based Imbalance degree (EID) using Eq. (8).

END WHILE

OUTPUT: Dataset X' which has a better-imbalanced ratio EID

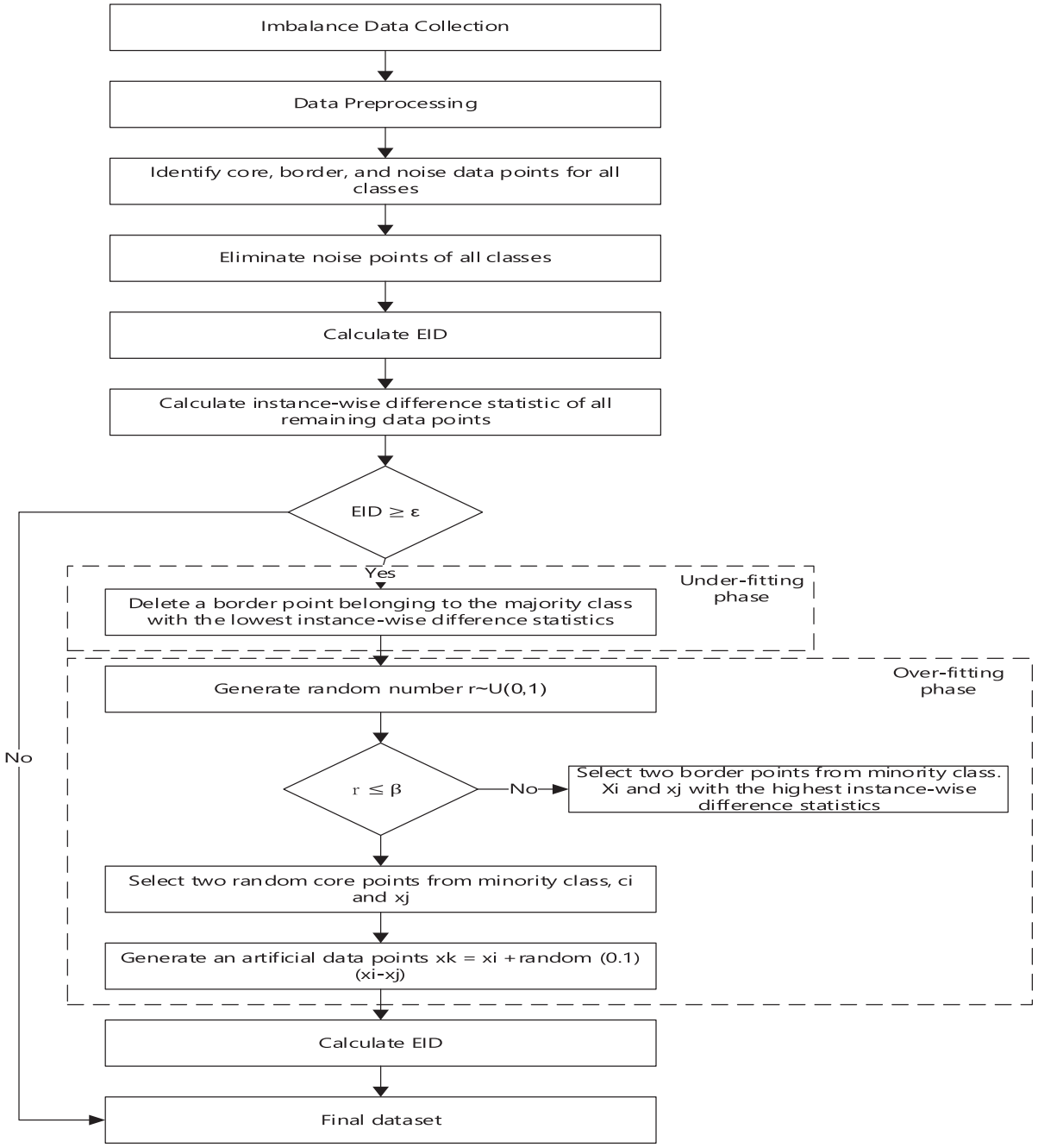

Figure 1: Entropy-based hybrid algorithm (IF-HA).

The dominant computational cost arises from recalculating EID and instance-wise difference statistics within the WHILE loop. Let N denote the number of instances, D, the number of features, k, the number of nearest neighbors, M, the number of classes, and T, the number of iterations until EID < ε. Scanning instance and feature statistics requires O(ND) per iteration. If nearest-neighbor computations are performed using brute-force search, this step has a worst-case complexity of O(N2D), while practical implementations with cached or approximate neighbors reduce this to O(NkD). Therefore, the total time complexity is O(TND) in optimized implementations or O(TN2D) in the worstcase scenario. Sampling operations (undersampling/oversampling) are linear in the number of selected instances and do not exceed the dominant entropy computation cost. Space complexity is O(ND + Nk) to store the dataset and neighbor statistics. This analysis demonstrates that IF-HA is computationally feasible for small- to medium-scale datasets, with approximate neighbor methods recommended for high-dimensional or large-scale applications.

The proposed method employs a two-phase hybrid sampling strategy that integrates entropy-based undersampling with entropy-based oversampling. Fig. 1 illustrates this process using a synthetic dataset.

In the first stage, entropy-based undersampling is applied to the majority class. Specifically, instances with low instance-wise entropy, reflecting limited contribution to class boundary definition, are removed. This step reduces redundant majority samples while preserving informative border points, thereby enhancing class separability. One potential limitation of entropy-guided undersampling is the possibility of removing structurally important majority border instances that contribute to the classifier decision boundary. Excessive removal may reduce majority class recall and oversimplify class separation. However, IF-HA attempts to mitigate this issue by removing only border instances with the minimum instance-wise difference statistic, which are assumed to contribute minimally to the overall entropy structure.

In the second stage, entropy-based oversampling generates synthetic minority class instances around the core and selected border regions. Unlike conventional oversampling, which interpolates blindly between minority points, our method uses local entropy values to determine where synthetic samples are most valuable for classification. For example, regions with high entropy, indicating uncertain decision boundaries, are prioritized for oversampling. This study evaluates synthetic minority class effectiveness indirectly through classification performance metrics.

The final distribution accomplishes two key objectives: first, it produces a cleaner majority cluster with reduced noise and redundancy, and second, it expands the minority class into a more representative space that reinforces decision boundaries. Collectively, these effects enhance the classifier's robustness and generalization capability.

3.4 Classification Algorithm

In this study, the effectiveness of the proposed IF-HA sampling method is evaluated using the k-nearest neighbors (KNN) classifier. KNN is a widely adopted benchmark in imbalance research because its performance is strongly influenced by class distribution, making it a suitable test bed for sampling strategies [28,51]. Using a simple and transparent classifier allows us to isolate the impact of the resampling method itself, without confounding effects from complex model architectures. Future work may extend the evaluation to ensemble and deep learning classifiers to further validate the generality of IF-HA.

4 Experiment and Discussion

This study conducts a series of experiments to evaluate the performance of the proposed IF-HA algorithm. The first experiment investigates the impact of different parameter settings and identifies the optimal configuration for the algorithm. Subsequently, the algorithm is applied to several benchmark

datasets, and its results are compared against those of existing methods. Finally, the proposed algorithm is validated through an implementation on a real-world case study.

4.1 Parameter Setting

In the proposed algorithm, three parameters must be pre-defined: k, ε, and β. Value k determines how many nearest neighbors are set for each data point. Nearest neighbors mean the closest points (based on distance) in the dataset. Furthermore, ε is used to control the final imbalanced quality, while β controls the oversampling procedure. Each parameter was varied while the others were fixed at their default values, and the average F1 Score across benchmark datasets was recorded. The results in Table 1 indicate that the algorithm is relatively stable within a specific value.

Table 1: Parameter setting.

Parameter	Tested Values	Best Value
k (Neighbours)	3, 4, and 5	4
ε (Imbalance control)	10^{-2}, 10^{-3}, and 10^{-4}	10^{-2}
β (Oversampling control)	0.01, 0.1, and 0.5	0.1

The oversampling control parameter β was evaluated using three candidate values: 0.01, 0.1, and 0.5. The value β = 0.1 was selected because it provided the best balance between generating sufficient minority class samples and avoiding excessive synthetic sample generation. A smaller value, β = 0.01, produces limited oversampling effects, resulting in lower improvement for minority class recognition and low F1 improvement. In contrast, β = 0.5 substantially increased oversampling intensity, which may introduce noisy or less representative synthetic samples and reduce generalization performance. Based on the ablation results, β = 0.1 achieved more stable classification performance across datasets and was therefore used as the final setting.

Following the determination of β, we investigated the robustness of the entropy-based stopping threshold ε, which controls the termination of the iterative sampling process by limiting the residual entropy-based imbalance degree (EID). Smaller ε values force additional iterations, producing more aggressive balancing, while larger values terminate the process earlier. Experimental observations indicate that ε = 10^{-2} provides the best trade-off between improving minority-class recognition and preventing overfitting. Extremely small thresholds ε = 10^{-4} tend to oversmooth the class boundary and increase computational cost, particularly in moderately imbalanced datasets, while larger thresholds may terminate before sufficient minority reinforcement is achieved.

Overall, these results suggest that IF-HA is relatively stable across moderate and severe imbalance conditions, with the entropy threshold serving as a practical convergence control mechanism rather than a dataset-specific tuning parameter. This combination of carefully tuned oversampling and threshold parameters ensures effective minority-class enhancement while maintaining robust generalization across datasets.

Although the sensitivity analysis provides useful evidence regarding the individual influence of k, ε and β, it should be noted that the current analysis follows a one-at-a-time tuning strategy, in which one parameter is varied while the others are fixed. This approach does not fully capture possible interaction effects among parameters. Consequently, different combinations of these parameters may influence the final resampled distribution and classifier performance. Therefore, the selected parameter values should be interpreted as empirically stable default settings rather than globally optimal values.

4.2 Evaluation Measures

Relying solely on accuracy can be misleading in imbalanced classification problems, as the majority class dominates. For instance, a classifier that predicts all instances as negative (majority class) can still achieve high accuracy despite completely failing to detect positive (minority class) instances. Therefore, more robust performance metrics are required.

This study employs four primary evaluation metrics: Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). Precision measures the proportion of correctly identified positive instances among all instances classified as positive. Recall quantifies the proportion of actual positives correctly classified. The F1-score, defined as the harmonic mean of Precision and Recall, provides a balanced measure emphasizing minority class detection.

In addition, this study also reports the Geometric Mean (G-Mean) in later experiments. Both metrics explicitly consider the trade-off between sensitivity and specificity, making them highly relevant for imbalanced classification tasks [55].

4.3 Experimental Result

This study evaluated the proposed algorithm using 5-fold cross-validation repeated across 30 independent runs. Results were averaged to minimize variance from random partitioning. The performance of the proposed IF-HA method was benchmarked against six sampling strategies: SMOTE BSMOTE (Borderline SMOTE), ADASYN, SMOTE-TLink (SMOTE-Tomek Links), SMOTE-NaN (Natural Neighbors)-DE (Differential Evolution), and SMOTE-TLNN-DEPSO. All experiments were implemented with k-nearest neighbors (KNN) as the base classifier, unless stated otherwise.

To assess the effectiveness of the proposed IF-HA method, this study compared classification performance before and after imbalance correction on 20 benchmark datasets. Performance was evaluated using accuracy, precision, recall, and F1-score, with an average of over 30 independent runs as presented in Table 2.

Table 2: Comparison before and after the imbalance.

No.	Dataset	Imbalanced	Accuracy	Precision	Recall	F1
1	Glass1 (IR 1.82)	Before	0.7856	0.7216	0.6475	0.677
		After	0.7966	0.8006	0.5705	0.6494
		Sig (2-tailed)	<0.001	0.011	0.072	0.416
2	Wisconsin (IR 1.86)	Before	0.9751	0.9843	0.9774	0.9807
		After	0.9738	0.9782	0.9819	0.9799
		Sig (2-tailed)	<0.001	0.042	<0.001	<0.001
3	Pima (IR 1.87)	Before	0.7213	0.7658	0.824	0.7936
		After	0.7588	0.9128	0.697	0.7889
		Sig (2-tailed)	<0.001	<0.001	<0.001	0.214
4	Glass0 (IR 2.06)	Before	0.7616	0.6225	0.6714	0.6418
		After	0.8171	0.7092	0.769	0.7222
		Sig (2-tailed)	<0.001	0.003	0.003	<0.001
5	Yeast1 (IR 2.46)	Before	0.7406	0.7911	0.8635	0.8256
		After	0.794	0.9274	0.7706	0.8409
		Sig (2-tailed)	<0.001	<0.001	<0.001	0.004

(Continued)

Table 2 (continued)

No.	Dataset	Imbalanced	Accuracy	Precision	Recall	F1
6	Haberman (IR 2.78)	Before	0.7025	0.7558	0.8799	0.8129
		After	0.7842	0.8816	0.8185	0.8478
		Sig (2-tailed)	<0.001	<0.001	<0.001	<0.001
7	Vehicle1 (IR 2.9)	Before	0.7187	0.7963	0.8362	0.8154
		After	0.759	0.9253	0.7355	0.8193
		Sig (2-tailed)	<0.001	<0.001	<0.001	0.322
8	Vehicle3 (IR 2.99)	Before	0.7494	0.5183	0.2877	0.3621
		After	0.7941	0.7296	0.2802	0.4006
		Sig (2-tailed)	<0.001	<0.001	0.636	0.023
9	Glass-0-1-2-3_Vs_4-5-6 (IR 3.2)	Before	0.905	0.8628	0.8287	0.8424
		After	0.8916	0.911	0.7147	0.7921
		Sig (2-tailed)	0.016	<0.001	<0.001	<0.001
10	Ecoli2 (IR 5.46)	Before	0.9583	0.9822	0.9684	0.9751
		After	0.9271	0.9823	0.9309	0.9555
		Sig (2-tailed)	<0.001	0.943	<0.001	<0.001
11	Glass6 (IR 6.38)	Before	0.9487	0.8695	0.7666	0.8011
		After	0.9627	0.96	0.7666	0.8472
		Sig (2-tailed)	<0.001	<0.001	NA	<0.001
12	Yeast3 (IR 8.1)	Before	0.9467	0.8119	0.675	0.7362
		After	0.955	0.8735	0.6926	0.7703
		Sig (2-tailed)	<0.001	<0.001	0.172	<0.001
13	Ecoli3 (IR 8.6)	Before	0.9285	0.74286	0.6	0.6263
		After	0.9299	0.6599	0.5143	0.5106
		Sig (2-tailed)	0.723	0.38	0.177	0.015
14	Page-Blocks0 (IR 8.79)	Before	0.9581	0.9676	0.9863	0.9769
		After	0.9607	0.9655	0.9916	0.9784
		Sig (2-tailed)	0.004	<0.001	<0.001	0.002
15	Biodegradation (IR 1.96)	Before	0.8159	0.7122	0.769	0.7379
		After	0.7984	0.6445	0.8985	0.7501
		Sig (2-tailed)	0.008	<0.001	<0.001	0.101
16	Indian Liver (IR 2.49)	Before	0.67	0.7503	0.8072	0.7774
		After	0.7396	0.9079	0.7072	0.7943
		Sig (2-tailed)	<0.001	<0.001	<0.001	<0.001
17	Liver (IR 2.49)	Before	0.683	0.6551	0.5345	0.5864
		After	0.7122	0.6291	0.8048	0.6978
		Sig (2-tailed)	0.068	0.3	<0.001	<0.001

(Continued)

Table 2 (continued)

No.	Dataset	Imbalanced	Accuracy	Precision	Recall	F1
18	Vertebral Column (IR 2.12)	Before	0.8349	0.8949	0.8611	0.8757
		After	0.8296	0.9711	0.772	0.8591
		Sig (2-tailed)	0.446	<0.001	<0.001	0.009
19	Wilt (IR 199.13)	Before	0.9797	0.9368	0.6692	0.7801
		After	0.9762	0.986	0.5654	0.7184
		Sig (2-tailed)	<0.001	<0.001	<0.001	<0.001
20	Banknote (IR 1.24)	Before	1	1	1	1
		After	0.9961	0.9915	1	0.9957
		Sig (2-tailed)	<0.001	<0.001	NA	<0.001

Overall, the performance of the IF-HA method is significantly improved for minority class detection, where it achieved substantial and statistically significant improvements in the F1-score. For instance, in the Glass0 dataset, the F1-score rose from 0.6418 to 0.7222 with a significance level of $p < 0.001$. Similar success was observed in the Yeast1 and Haberman datasets, where F1-scores increased from 0.8256 to 0.8409 ($p = 0.004$) and 0.8129 to 0.8478 ($p < 0.001$), respectively. Most remarkably, the Liver dataset saw a major performance leap, with the F1-score improving significantly from 0.5864 to 0.6978 ($p < 0.001$), underscoring the method's effectiveness in enhancing classification outcomes for these specific imbalanced scenarios.

In datasets like Pima, IF-HA transformed the classifier from a high-recall/low-precision model to a high-precision specialist, with a precision of 0.76 to 0.91, which is vital in medical contexts where false positives are costly. In extreme cases like Wilt or Ecoli3, the entropy-based filtering may be too aggressive, removing instances that the base classifier relies on for the F1-metric.

4.4 Statistical Test

To evaluate whether the performance changes produced by IF-HA were statistically significant, this study employed a paired sample *t*-test using Sig. (2-tailed) values reported in Table 2. The paired *t*-test is appropriate because it examines whether the mean difference between two related conditions classifier performance before and after applying IF-HA on the same dataset is significantly different from zero. In this study, a significance level of 0.05 was used to determine whether the observed differences were statistically meaningful.

The results indicate that the effect of IF-HA varies across evaluation metrics and datasets, with the strongest evidence observed in the F1-score. In several datasets, IF-HA produced statistically significant improvements in F1-score, such as Glass0 (from 0.6418 to 0.7222, $p < 0.001$), Yeast1 (from 0.8256 to 0.8409, $p = 0.004$), Haberman (from 0.8129 to 0.8478, $p < 0.001$), Glass6 (from 0.8011 to 0.8472, $p < 0.001$), Yeast3 (from 0.7362 to 0.7703, $p < 0.001$), Indian Liver (from 0.7774 to 0.7943, $p < 0.001$), and Liver (from 0.5864 to 0.6978, $p < 0.001$). These results suggest that IF-HA is effective in improving the balance between Precision and Recall, which is particularly important in imbalanced classification problems.

However, the improvements are not uniform across all metrics. Although Precision increased significantly in many datasets, Recall did not always improve and in some cases decreased significantly. For instance, in the Indian Liver dataset, Recall declined from 0.8072 to 0.7072 ($p < 0.001$), while Precision increased from 0.7503 to 0.9079 ($p < 0.001$), resulting in a higher F1-score overall. A similar trade-off can

be observed in Yeast1 and Vehicle1, where Recall decreased after resampling while Precision increased. This indicates that IF-HA tends to make the classifier more selective, reducing false positives in some cases, but not always increasing sensitivity to the minority class.

Overall, the statistical evidence supports that IF-HA provides significant improvement mainly in F1-score and, in many cases, Precision, while its impact on Recall and Accuracy is more dataset-dependent. Therefore, the main strength of IF-HA lies in enhancing the overall balance of minority-class prediction rather than consistently improving every performance metric. These findings demonstrate that IF-HA is a robust resampling approach for many imbalanced datasets, although its effectiveness remains influenced by the specific distributional characteristics of each dataset.

4.5 Comparison with Other Algorithms

This study has compared the performance of the proposed IF-HA approach with SMOTE [28], Borderline SMOTE (BSMOTE) [29], ADASYN [33], SMOTE-TLink [51], SMOTE-NaN-DE [58], and SMOTE-TLNN-DEPSO [58]. This study has used the K Nearest Neighbor (KNN) algorithm [58] as the test classifier. Standard parameters are used for all the approaches mentioned in their standard versions. Tables 3–5 describe the accuracies, F1-measure, and G-mean, respectively, of SMOTE, Borderline SMOTE (BSMOTE), ADASYN, SMOTE-TLink, SMOTE-NaN-DE, SMOTE-TLNN-DEPSO, and the proposed IF-HA.

Table 3: Accuracy of datasets (%).

Dataset	SMOTE	BSMOTE	ADASYN	SMOTE-TLink	SMOTE-NaN-DE	SMOTE-TLNN-DEPSO	IF-HA
Biodegradation	84.54	87.56	71.63	89.21	93.54	94.12	79.84
Indian Liver	51.22	51.35	53.43	54.25	64.82	65.34	73.96
Liver	66.23	67.32	64.34	67.56	68.32	73.23	71.22
Vertebral Column	68.63	72.04	69.43	72.11	73.09	76.43	82.95
Wilt	78.37	76.25	76.25	79.9	82.14	84.32	97.62
Banknote	91.67	91.79	94.14	96.17	97.24	99.67	99.61
Ecoli2	89.3	95.6	96.3	96.1	92.4	97.1	92.71
Haberman	86.32	91.82	90.96	91.93	92	92.91	78.42

Table 4: F-1 measures the datasets (%).

Dataset	SMOTE	BSMOTE	ADASYN	SMOTE-TLink	SMOTE-NaN-DE	SMOTE-TLNN-DEPSO	IF-HA
Biodegradation	72.59	68.62	71.63	68.96	71.97	73.54	75.02
Indian Liver	45.09	49.18	45.24	50.58	51.16	55.34	79.42
Liver	64.69	61.78	64.34	64.48	67.21	69.21	69.78
Vertebral Column	68.63	72.04	69.43	72.11	73.09	76.43	85.92
Wilt	77.1	75.52	75.52	78.8	81.24	83.22	71.85
Banknote	90	92.9	93.6	95.7	96.4	99.3	99.56
Ecoli2	92.9	90	93.6	95.7	96.4	99.3	95.57
Haberman	42.08	46.07	39.17	44.65	46.18	52.34	84.79

Table 5: G-mean measure of the datasets (%).

Dataset	SMOTE	BSMOTE	ADASYN	SMOTE-TLink	SMOTE-NaN-DE	SMOTE-TLNN-DEPSO	IF-HA
Biodegradation	78.84	76.34	77.88	77.17	79.15	81.67	76.10
Indian Liver	60.39	63.31	60.21	63.7	64.62	67.13	80.13
Liver	55.91	59.87	64.2	62.45	68.25	71.21	71.15
Vertebral Column	79.66	82.69	80.19	81.08	83.98	85.43	86.58
Wilt	82.23	79.64	82.82	81.2	82.94	84.32	74.66
Banknote	89.6	94	95.5	96.4	96.8	97.4	99.57
Ecoli2	80.1	82.7	83.4	83.8	83.5	84.7	95.63
Haberman	58.55	60.3	55.94	60.47	61	63.45	84.95

IF-HA demonstrates highly competitive results when considering accuracy, particularly on the wilt, BankNote, and vertebral column datasets. These outcomes are notably higher than those achieved by conventional over-sampling methods such as SMOTE or ADASYN, as well as hybrid approaches like SMOTE-TLNN-DEPSO. For example, on the BankNote dataset, IF-HA achieved nearly perfect accuracy of 99.61%, remaining highly competitive with the best-performing baseline 99.67%. On more challenging datasets, such as Indian Liver and Liver, IF-HA also achieved strong accuracy levels 73.96% and 71.22%, respectively, outperforming traditional methods that often struggled to exceed 65%. These results confirm that IF-HA is not only effective at improving minority class representation but also succeeds in maintaining high predictive accuracy across diverse domains.

The F1-score results further emphasize the superiority of IF-HA, particularly in datasets characterized by higher imbalance ratios. For instance, on Indian Liver and Liver, IF-HA achieved F1-scores of 79.42% and 69.78%, respectively, representing significant improvements over SMOTE and ADASYN, which typically produced values below 65%. On the Banknote dataset, IF-HA again delivered the best outcome with 99.56%, exceeding the strongest competing baseline (99.3% from SMOTE-TLNN-DEPSO). Similarly, in the vertebral column and haberman, IF-HA obtained 85.92% and 84.79%, respectively, outperforming the baselines. These findings demonstrate that IF-HA improves precision and recall balance and effectively strengthens minority class detection, the central challenge in imbalanced classification tasks.

The G-Mean metric, which balances sensitivity and specificity, further validates IF-HA's robustness. On datasets such as vertebral column 86.58%, ecoli2 95.63%, and Haberman 84.95%, IF-HA achieved higher G-Mean values compared to the baseline methods. In the Indian liver, IF-HA reached a G-Mean of 80.13%, far exceeding traditional sampling methods, mostly below 65%. These improvements indicate that IF-HA can simultaneously increase minority class recognition while minimizing false positives from the majority class.

4.6 Case Study

The proposed algorithm is also applied to a real-world dataset, collected from patient data of tuberculosis (TB) cases in Indonesia. The increasing prevalence of TB, particularly during the COVID-19 pandemic, has exacerbated public health challenges and imposed significant economic burdens on communities. Healthcare institutions have been striving to address this situation by providing diagnostic, treatment, and preventive services to patients across various demographics. These services are critical for managing TB cases effectively and reducing its transmission within communities.

A significant challenge healthcare institutions face is ensuring patients adhere to their prescribed treatment regimens. Non-adherence to TB treatment, often due to negligence, lack of awareness, or socio-economic barriers, leads to higher rates of drug resistance and prolonged infectiousness. This strains healthcare systems and undermines efforts to control the disease. Therefore, identifying patients likely to adhere to their treatment regimens is essential. This can be achieved through classification methods.

The TB dataset contains 480 instances with an imbalance ratio of 1:2.43 between minority and majority classes, with 140 and 340 each. The dataset includes both categorical and numerical attributes. TB dataset has nine features: body mass index, sex, age, education, economic status, smoking habits, household contacts, family size, and BCG immunization. Missing values and duplicated records were removed during preprocessing. The dataset was evaluated using the same 5-fold cross-validation repeated over 30 runs used in the benchmark experiments to ensure consistency and robustness. Table 6 shows the Description of case study dataset.

Table 6: Description of case study dataset.

Information	Description	Attribute
Body Mass Index	A calculated value derived from weight and height (kg/m^2)	Categorical (underweight/normal/overweight)
Sex	The patient's biological sex is assigned at birth	Categorical (Male/Female)
Age	The age of the patient	Categorical (<31 years/31–43 years/44–57 years/>57 years)
Education	The highest formal educational level attained by the patient	Categorical (Elementary school/Junior High School/Senior High School/Associate Degree/Undergraduate/Master)
Economic Status	Socioeconomic classification based on income, employment, and living conditions	Categorical (Poor/Not Poor)
Smoking Habit	Whether the patient is a smoker or not	Categorical (Yes/No)
Household Contact	Close contact with individuals diagnosed with active tuberculosis within the household environment	Categorical (Positive/Negative)
Family Size	The total number of individuals residing in the patient's household	Numeric
BCG Vaccine	Whether the patient has had the BCG vaccine or not	Categorical (Yes/No)
Tuberculosis (TB)	Diagnosis of active tuberculosis infection	Categorical (Yes/No)

For the Tuberculosis (TB) dataset case study, the proposed IF-HA method was evaluated using three classification algorithms: k-nearest neighbors (KNN), Random Forest, and XGBoost (XGB). KNN was selected because of its simplicity, transparency, and sensitivity to class distribution, while Random Forest

and XGBoost were included to examine whether the proposed resampling method remains effective when applied to stronger ensemble-based classifiers commonly used in practical medical prediction tasks. Table 7 presents the classification performance before and after applying IF-HA across the three classifiers. The results show that IF-HA consistently improved Accuracy, Precision, Recall, and F1-score for all classifiers. These results indicate that IF-HA enhances minority-class representation and improves predictive performance across both distance-based and ensemble-based classifiers. Therefore, the observed improvement is not specific to KNN but reflects the ability of IF-HA to improve the underlying training data distribution for TB classification.

Table 7: Improvement results for TB dataset (%).

Classifier Method	TB Dataset	Accuracy	Precision	Recall	F1
KNN	Before	68.54	70.95	73.06	70.58
	After	83.47	88.97	85.48	86.87
Random forest	Before	72.56	59.05	57.11	54.39
	After	94.34	91.61	94.783	92.66
XGB	Before	65.83	58.67	55.71	50.18
	After	93.64	91.53	93.44	91.83

From a healthcare perspective, these improvements are meaningful. Higher recall indicates fewer high-risk patients are overlooked, while higher precision reduces false alarms. This balance is crucial for resource-constrained healthcare systems, as it allows targeted interventions for patients most likely to discontinue treatment. Consequently, IF-HA enhances predictive accuracy, improves TB management, and reduces disease transmission in affected communities.

5 Conclusions

This study introduces the Information Filtered Hybrid Algorithm (IF-HA), a novel entropy-based hybrid sampling method designed to address the persistent challenge of imbalanced data classification. By leveraging information theory, the proposed method quantifies the importance of individual instances and balances class distributions through an iterative process that combines under-sampling of the majority class and over-sampling of the minority class. This process strategically eliminates noise and enhances the representativeness of synthetic data, leading to more robust and fair classification models. The experimental results in Table 3 demonstrate that the IF-HA method achieves a statistically significant impact on classification performance across the majority of tested datasets, $p < 0.05$. Specifically, for 12 of the 20 datasets, the proposed method either maintained or significantly improved the F1-score. Most notably, in the Liver and Glass0 datasets, IF-HA produced substantial gains in F1-performance, supported by p-values < 0.001. While certain datasets showed a trade-off between Recall and Precision, the overall trend confirms that IF-HA effectively refines the minority class boundary, leading to higher Precision and stabilized Accuracy

Extensive evaluations across benchmark datasets and a real-world healthcare case demonstrate that IF-HA consistently outperforms traditional sampling methods in improving F1-scores and AUC values. The entropy-based imbalance degree metric proves to be an effective criterion for guiding resampling operations, ensuring that the final dataset maintains both class balance and data quality. Moreover, IF-HA integrates well with a range of machine learning classifiers, showcasing its adaptability and scalability.

These findings highlight the potential of combining data-level strategies with information measures to overcome the limitations of existing resampling approaches. Future work should explore extensions of IF-HA to multi-class classification for high dimensional domains such as medical imaging and dynamic online learning settings. The integration of domain knowledge or meta-learning to guide parameter selection could further enhance the algorithm's performance in real-world applications.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design: Ren-Jieh Kuo, Muhammad Rizki, Ferani Eva Zulvia; Data collection: Eddy Roflin; Analysis and interpretation of methods: Ren-Jieh Kuo, Muhammad Rizki, Ferani Eva Zulvia; Draft manuscript preparation: Ren-Jieh Kuo, Muhammad Rizki; Review and editing: Ren-Jieh Kuo, Muhammad Rizki, Ferani Eva Zulvia. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets are publicly available on the KEEL (Knowledge Extraction Based on Evolutionary Learning), the dataset can be accessed via: https://sci2s.ugr.es/keel/imbalanced.php#subA and the University of California, Irvine (UCI) repositories, which are recommended for imbalanced datasets in classification studies the dataset can be accessed via: https://archive.ics.uci.edu/datasets. The tuberculosis dataset was collected through a survey involving human respondents and is only available for the current study due to ethical and confidentiality considerations.

Ethics Approval: The study protocol, informed consent procedure, and data collection process were reviewed and approved by the Ethics Committee of the Faculty of Medicine, Sriwijaya University (Approval No. 0068/UN9/SB.SU/2020). All participants provided written informed consent.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Carvalho M, Pinho AJ, Brás S. Resampling approaches to handle class imbalance: a review from a data perspective. J Big Data. 2025;12(1):71. doi:10.1186/s40537-025-01119-4.
2. Xie J, Sun L, Zhao YF. On the data quality and imbalance in machine learning-based design and manufacturing—a systematic review. Engineering. 2025;45:105–31. doi:10.1016/j.eng.2024.04.024.
3. Zhang Y, Muniyandi RC, Qamar F. A review of deep learning applications in intrusion detection systems: overcoming challenges in spatiotemporal feature extraction and data imbalance. Appl Sci. 2025;15(3):1552. doi:10.3390/app15031552.
4. Wang S. A hybrid SMOTE and trans-CWGAN for data imbalance in real operational AHU AFDD: a case study of an auditorium building. Energy Build. 2025;348:116447. doi:10.1016/j.enbuild.2025.116447.
5. Ando S, Huang CY. Deep over-sampling framework for classifying imbalanced data. In: Machine learning and knowledge discovery in databases. Cham, Switzerland: Springer; 2017. p. 770–85. doi:10.1007/978-3-319-71249-9_46.
6. Cao P, Zhao D, Zaiane O. An optimized cost-sensitive SVM for imbalanced data learning. In: Advances in knowledge discovery and data mining. Berlin/Heidelberg, Germany: Springer; 2013. p. 280–92. doi:10.1007/978-3-642-37456-2_24.
7. Krawczyk B, Woźniak M, Schaefer G. Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl Soft Comput. 2014;14:554–62. doi:10.1016/j.asoc.2013.08.014.
8. Zhang C, Tan KC, Li H, Hong GS. A cost-sensitive deep belief network for imbalanced classification. IEEE Trans Neural Netw Learn Syst. 2019;30(1):109–22. doi:10.1109/tnnls.2018.2832648.

9. Shawon MTR, Shahariar Shibli GM, Ahmed F, Joy SKS. Explainable cost-sensitive deep neural networks for brain tumor detection from brain MRI images considering data imbalance. Multimed Tools Appl. 2025;84(35):43615–42. doi:10.1007/s11042-025-20842-x.
10. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C Appl Rev. 2012;42(4):463–84. doi:10.1109/TSMCC.2011.2161285.
11. Khoshgoftaar TM, Van Hulse J, Napolitano A. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst Man Cybern Part A Syst Hum. 2011;41(3):552–68. doi:10.1109/TSMCA.2010.2084081.
12. López V, Fernández A, Moreno-Torres JG, Herrera F. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Syst Appl. 2012;39(7):6585–608. doi:10.1016/j.eswa.2011.12.043.
13. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32. doi:10.1007/s13748-016-0094-0.
14. Makki S, Assaghir Z, Taher Y, Haque R, Hacid MS, Zeineddine H. An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access. 2019;7:93010–22. doi:10.1109/ACCESS.2019.2927266.
15. Liu K, Bao C, Liu S. Semi-supervised medical image classification based on sample intrinsic similarity using canonical correlation analysis. Comput Mater Contin. 2025;82(3):4451–68. doi:10.32604/cmc.2024.059053.
16. Chui KT, Arya V, Gupta BB, Torres-Ruiz M, Attar RW. A convolutional neural network-based deep support vector machine for Parkinson's disease detection with small-scale and imbalanced datasets. Comput Mater Contin. 2026;86(1):1–23. doi:10.32604/cmc.2025.068842.
17. Ding H, Huang N, Wu Y, Cui X. LEGAN: addressing intraclass imbalance in GAN-based medical image augmentation for improved imbalanced data classification. IEEE Trans Instrum Meas. 2024;73:2517914. doi:10.1109/TIM.2024.3396853.
18. Yang Y, Tang X, Liu Z, Cheng J, Fang H, Zhang C. Diff-IDS: a network intrusion detection model based on diffusion model for imbalanced data samples. Comput Mater Contin. 2025;82(3):4389–408. doi:10.32604/cmc.2025.060357.
19. Bagui S, Li K. Resampling imbalanced data for network intrusion detection datasets. J Big Data. 2021;8(1):6. doi:10.1186/s40537-020-00390-x.
20. Abdelkhalek A, Mashaly M. Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning. J Supercomput. 2023;79(10):10611–44. doi:10.1007/s11227-023-05073-x.
21. Sambasivam G, Opiyo GD. A predictive machine learning application in agriculture: cassava disease detection and classification with imbalanced dataset using convolutional neural networks. Egypt Inform J. 2021;22(1):27–34. doi:10.1016/j.eij.2020.02.007.
22. Miftahushudur T, Sahin HM, Grieve B, Yin H. A survey of methods for addressing imbalance data problems in agriculture applications. Remote Sens. 2025;17(3):454. doi:10.3390/rs17030454.
23. Tian J, Jiang Y, Zhang J, Luo H, Yin S. A novel data augmentation approach to fault diagnosis with class-imbalance problem. Reliab Eng Syst Saf. 2024;243(4):109832. doi:10.1016/j.ress.2023.109832.
24. Chen J, Yan Z, Lin C, Yao B, Ge H. Aero-engine high speed bearing fault diagnosis for data imbalance: a sample enhanced diagnostic method based on pre-training WGAN-GP. Measurement. 2023;213:112709. doi:10.1016/j.measurement.2023.112709.
25. Li Y, Yang Z, Xing L, Yuan C, Liu F, Wu D, et al. Crash injury severity prediction considering data imbalance: a Wasserstein generative adversarial network with gradient penalty approach. Accid Anal Prev. 2023;192(2):107271. doi:10.1016/j.aap.2023.107271.
26. Guo J, Wu H, Chen X, Lin W. Adaptive SV-borderline SMOTE-SVM algorithm for imbalanced data classification. Appl Soft Comput. 2024;150:110986. doi:10.1016/j.asoc.2023.110986.
27. Vairetti C, Assadi JL, Maldonado S. Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification. Expert Syst Appl. 2024;246(2):123149. doi:10.1016/j.eswa.2024.123149.

28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Jair. 2002;16:321–57. doi:10.1613/jair.953.
29. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Advances in intelligent computing. Berlin/Heidelberg, Germany: Springer; 2005. p. 878–87. doi:10.1007/11538059_91.
30. Al Majzoub H, Elgedawy I. AB-SMOTE: an affinitive borderline SMOTE approach for imbalanced data binary classification. Int J Mach Learn Comput. 2020;10(1):31–7. doi:10.18178/ijmlc.2020.10.1.894.
31. Al Majzoub H, Elgedawy I, Akaydın Ö, Köse Ulukök M. HCAB-SMOTE: a hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification. Arab J Sci Eng. 2020;45(4):3205–22. doi:10.1007/s13369-019-04336-1.
32. Hussein AS, Li T, Yohannese CW, Bashir K. A-SMOTE: a new preprocessing approach for highly imbalanced datasets by improving SMOTE. Int J Comput Intell Syst. 2019;12(2):1412–22. doi:10.2991/ijcis.d.191114.002.
33. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008 Jun 1–8; Hong Kong, China. p. 1322–8. doi:10.1109/IJCNN.2008.4633969.
34. Fu GH, Wang JB, Lin W. An adaptive loss backward feature elimination method for class-imbalanced and mixed-type data in medical diagnosis. Chemom Intell Lab Syst. 2023;236:104809. doi:10.1016/j.chemolab.2023.104809.
35. Mienye ID, Sun Y. A deep learning ensemble with data resampling for credit card fraud detection. IEEE Access. 2023;11(6):30628–38. doi:10.1109/ACCESS.2023.3262020.
36. Hart P. The condensed nearest neighbor rule (Corresp.). IEEE Trans Inform Theory. 1968;14(3):515–6. doi:10.1109/tit.1968.1054155.
37. Laurikkala J. Improving identification of difficult small classes by balancing class distribution. In: Artificial intelligence in medicine. Berlin/Heidelberg, Germany: Springer; 2001. p. 63–6. doi:10.1007/3-540-48229-6_9.
38. Devi D, Biswas SK, Purkayastha B. Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance. Pattern Recognit Lett. 2017;93(1):3–12. doi:10.1016/j.patrec.2016.10.006.
39. Anand A, Pugalenthi G, Fogel GB, Suganthan PN. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids. 2010;39(5):1385–91. doi:10.1007/s00726-010-0595-2.
40. Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl. 2009;36(3):5718–27. doi:10.1016/j.eswa.2008.06.108.
41. Chen C, Shyu ML. Clustering-based binary-class classification for imbalanced data sets. In: Proceedings of the 2011 IEEE International Conference on Information Reuse & Integration; 2011 Aug 3–5; Las Vegas, NV, USA. p. 384–9. doi:10.1109/IRI.2011.6009578.
42. Lin WC, Tsai CF, Hu YH, Jhang JS. Clustering-based undersampling in class-imbalanced data. Inf Sci. 2017;409–410(1):17–26. doi:10.1016/j.ins.2017.05.008.
43. Shobana G, Prakash Battula B. An under sampled k-means approach for handlingimbalanced data using diversified distribution. Int J Eng Technol. 2018;1(8):113–7. doi:10.14419/ijet.v7i1.8.9984.
44. Koziarski M. Radial-based undersampling for imbalanced data classification. Pattern Recognit. 2020;102(4):107262. doi:10.1016/j.patcog.2020.107262.
45. Xie X, Liu H, Zeng S, Lin L, Li W. A novel progressively undersampling method based on the density peaks sequence for imbalanced data. Knowl Based Syst. 2021;213:106689. doi:10.1016/j.knosys.2020.106689.
46. Ha J, Lee JS. A new under-sampling method using genetic algorithm for imbalanced data classification. In: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication; 2016 Jan 4–6; Danang, Vietnam. New York, NY, USA: The Association for Computing Machinery (ACM); 2016. p. 1–6. doi:10.1145/2857546.2857643.
47. Prachuabsupakij W. CLUS: a new hybrid sampling classification for imbalanced data. In: Proceedings of the 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE); 2015 Jul 22–24; Songkhla, Thailand. p. 281–6. doi:10.1109/JCSSE.2015.7219810.

48. Sáez JA, Luengo J, Stefanowski J, Herrera F. SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci. 2015;291(C):184–203. doi:10.1016/j.ins.2014.08.051.
49. Zhang X, Zhu C, Wu H, Liu Z, Xu Y. An imbalance compensation framework for background subtraction. IEEE Trans Multimed. 2017;19(11):2425–38. doi:10.1109/TMM.2017.2701645.
50. Li J, Fong S, Wong RK, Chu VW. Adaptive multi-objective swarm fusion for imbalanced data classification. Inf Fusion. 2018;39(1):1–24. doi:10.1016/j.inffus.2017.03.007.
51. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl. 2004;6(1):20–9. doi:10.1145/1007730.1007735.
52. Ramentol E, Caballero Y, Bello R, Herrera F. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst. 2012;33(2):245–65. doi:10.1007/s10115-011-0465-6.
53. Bruhn J, Lehmann LE, Röpcke H, Bouillon TW, Hoeft A. Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane. Anesthesiology. 2001;95(1):30–5. doi:10.1097/00000542-200107000-00010.
54. Zhu C, Wang Z. Entropy-based matrix learning machine for imbalanced data sets. Pattern Recognit Lett. 2017;88(14):72–80. doi:10.1016/j.patrec.2017.01.014.
55. Li L, He H, Li J. Entropy-based sampling approaches for multi-class imbalanced problems. IEEE Trans Knowl Data Eng. 2020;32(11):2159–70. doi:10.1109/TKDE.2019.2913859.
56. Chan CP, Yang JH, Chang WH. Entropy-based time-series financial distress model based on attribute selection and MetaCost methods for imbalance class. In: Proceedings of the 2023 3rd International Conference on Artificial Intelligence, Automation and Algorithms; 2023 Jul 21–23; Beijing, China. New York, NY, USA: The Association for Computing Machinery (ACM); 2023. p. 140–51. doi:10.1145/3611450.3611471.
57. Li S, Li L, Yan J, He H. SDE: a novel clustering framework based on sparsity-density entropy. IEEE Trans Knowl Data Eng. 2018;30(8):1575–87. doi:10.1109/TKDE.2018.2792021.
58. Li J, Zhu Q, Wu Q, Zhang Z, Gong Y, He Z, et al. SMOTE-NaN-DE: addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution. Knowl Based Syst. 2021;223(3):107056. doi:10.1016/j.knosys.2021.107056.