Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Feature Representation Matters: Preprocessing Effects on One- Class Classification for Class-Imbalanced Medical Data

Min-Wei Huang
Kaohsiung Medical University
Chih-Fong Tsai
National Central University
Wei-Chao Lin
vicl in@gap.cgu.edu.tw
Chang Gung University

Research Article

Keywords: class imbalance, feature selection, instance selection, machine learning, one-class classification

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11091745/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Class-imbalanced medical datasets present significant challenges for traditional binary classifiers, particularly when minority-class instances represent clinically critical cases. One-class classification (OCC) has emerged as an effective alternative for addressing such imbalance by learning decision boundaries from only the target class. However, while previous studies have primarily focused on OCC algorithm design, the impact of data preprocessing on OCC performance remains insufficiently explored. In particular, how feature representation changes induced by feature selection and instance selection affect OCC performance has not been systematically investigated. Therefore, this study examines the effects of different preprocessing strategies on OCC-based medical classification using seven hard-to-classify medical datasets selected from the KEEL repository. Three OCC algorithms, namely one-class support vector machine (OCSVM), isolation forest (IF), and local outlier factor (LOF), are evaluated under multiple preprocessing procedures, including feature selection (FS), instance selection (IS), and their combinations in two different orders: FS followed by IS (FSIS) and IS followed by FS (ISFS). Experimental results show that preprocessing significantly improves OCC performance in terms of both AUC and recall compared with baseline models. Among individual preprocessing methods, CNN-based instance selection yields the best overall performance, particularly when combined with LOF, while CART-based feature selection performs better for highly imbalanced datasets. For combined preprocessing strategies, CART-ENN in the FSIS framework and ENN-CART in the ISFS framework achieve the best results, demonstrating that combining feature and instance selection further enhances detection performance. The findings also reveal that preprocessing effectiveness is influenced by class imbalance severity, indicating that adaptive preprocessing strategies are essential for robust OCC-based medical prediction systems. These results provide practical insights into designing effective preprocessing pipelines for one-class classification in class-imbalanced medical applications.

1. Introduction

Many real-world medical datasets suffer from class imbalance, where one class, typically the minority class, contains significantly fewer instances than the other. In breast cancer prediction, for example, malignant tumor cases generally occur far less frequently than benign cases. Such imbalance, characterized by varying ratios between minority and majority classes, poses a substantial challenge for traditional two-class classification algorithms. These algorithms often become biased toward the majority class, resulting in suboptimal performance, particularly in detecting minority class instances, which are often of greater clinical importance (Araf et al., 2024; Aubaidan et al., 2025; Salmi et al., 2024).

In contrast to conventional two-class approaches, one-class classification (OCC) algorithms are trained using examples from only a single class. OCC methods have been proposed for handling class-imbalanced datasets in domains such as anomaly detection, novelty detection, and outlier detection (Baro and Borah, 2025; Hayashi and Fujita, 2022; Hayashi et al., 2026; Leevy et al., 2025; Marques et al., 2023; Seliya et al., 2021).

Moreover, OCC methods have been applied to a wide range of medical datasets and problems, including the Medical Information Mart for Intensive Care (MIMIC) dataset (Ouzar et al., 2025), noninvasive isocitrate dehydrogenase (IDH) mutation detection (Jian et al., 2025), follicular lymphoma classification (Somaratne et al., 2023), quality assurance for organ-at-risk contour delineation (Zhao et al., 2025), melanoma recognition (Yu et al., 2023), cognitive impairment assessment (Guimaraes et al., 2025), and diabetes infection prediction (Woldaregay et al., 2020), among others.

However, although prior studies have demonstrated the effectiveness of one-class classification (OCC) methods across a variety of binary medical classification tasks, the impact of data preprocessing on OCC performance remains insufficiently explored. In particular, preprocessing strategies that remove unrepresentative features or instances can fundamentally alter the data distribution used to model the target class. While feature selection, which reduces feature dimensionality (Dhal and Azad, 2022; Moslemi, 2023; Theng and Bhoyar, 2023) and instance selection, which filters noisy samples or outliers (Fu et al., 2024; Huang et al., 2026; Qin et al., 2024) have been widely investigated in conventional supervised learning, their influence on OCC-based medical classification has received limited attention, leaving an important methodological gap in the current literature.

From a representation-learning perspective, feature representation plays a critical role in determining classification performance, especially in OCC settings where decision boundaries are inferred solely from a single-class distribution. Different representations can reshape the geometric structure of the data space, affecting class compactness, separability, and sensitivity to noise. High-dimensional representations may include redundant or irrelevant features that distort distance- or density-based OCC models, resulting in overly complex boundaries and reduced generalization. Conversely, representations refined through feature selection may emphasize informative attributes, thereby improving model stability and reducing overfitting.

In addition to feature-level transformations, instance-level modifications such as removing noisy or borderline samples can further influence the learned class representation by reducing variance and improving distributional consistency. Nevertheless, excessive filtering may eliminate clinically meaningful variability, potentially reducing model robustness and limiting generalizability. Consequently, feature representation directly affects not only predictive accuracy but also clinically relevant performance measures such as sensitivity and specificity. Understanding how preprocessing-induced representation changes interact with OCC behavior is therefore essential for developing reliable medical classification systems under class-imbalanced conditions.

Therefore, the objective of this study is to evaluate the performance of different one-class classifiers trained on varying feature representations obtained through feature selection and instance selection. In addition, because medical datasets often exhibit different degrees of class imbalance, this study also examines the relationship between feature representations and class imbalance ratios.

This paper makes the following key contributions:

1. Addressing an overlooked research gap in OCC-based medical classification. Although one-class classification (OCC) has been widely adopted for class-imbalanced medical problems, existing studies primarily focus on algorithmic design while largely overlooking how data preprocessing alters the learned class representation. This work provides one of the first dedicated investigations into the impact of preprocessing-induced feature representation changes on OCC performance in medical datasets.
2. A unified and systematic evaluation of feature representation effects on OCC. We establish a comprehensive experimental framework that systematically evaluates multiple OCC algorithms under feature representations generated through feature selection and instance selection. This framework enables direct analysis of how preprocessing reshapes the decision boundary learned from a single-class distribution, an aspect rarely examined in prior OCC research.
3. Novel investigation of the interaction between feature representation and class imbalance severity. Beyond conventional performance comparisons, this study explicitly examines how varying class imbalance ratios influence OCC behavior under different preprocessing strategies. This analysis reveals previously underexplored relationships between data imbalance, representation quality, and one-class learning effectiveness.
4. Practical design insights for robust OCC pipelines in medical AI. The findings provide evidence-based guidelines for integrating preprocessing strategies into OCC workflows, offering practical value for designing more reliable and robust medical classification systems under severe class imbalance conditions.

The remainder of this paper is organized as follows. Section 2 provides an overview of feature selection, instance selection, and one-class classification methods. Section 3 describes the research methodology, including the experimental framework used to generate different feature representations, the integration of various OCC methods, and the experimental design. Section 4 presents the experimental results, and Section 5 concludes the paper.

2. Literature Review

2.1 Feature Selection

Feature selection is one of the most widely used techniques in data preprocessing. By removing redundant features from a dataset, feature selection reduces the dataset's dimensionality and computational complexity, thereby improving the classification accuracy of subsequent classifiers and reducing computational time. This technique has been shown to produce positive effects in data mining and machine learning tasks, particularly when dealing with high-dimensional datasets (Dhal and Azad, 2022; Moslemi, 2023; Theng and Bhoyar, 2023).

The feature selection process can generally be divided into four steps. The first step, Generation, produces candidate feature subsets for evaluation. The second step, Evaluation, assesses each candidate subset to determine whether it performs better than the current best subset; if so, the best

subset is updated accordingly. The third step, the Stopping Criterion, defines the conditions under which the feature selection process terminates. The final step, Validation, verifies the selected feature subset and evaluates its effectiveness through various tests and comparisons.

In general, feature selection techniques can be categorized into three types: filter, wrapper, and embedded methods. Filter methods determine thresholds based on feature relevance and divergence, and features that satisfy these thresholds are considered important attributes. These methods do not incorporate classifiers into the selection process; instead, they evaluate each feature independently without considering interactions among features. As a result, filter-based methods generally offer faster computational speed compared with the other two categories of feature selection techniques.

Wrapper methods evaluate the importance of feature subsets by considering potential interactions among features, thereby identifying the optimal subset that yields the best performance. Unlike filter methods, wrapper approaches place greater emphasis on the actual impact of selected features on model performance. Consequently, their computational cost is higher than that of the other two categories of feature selection techniques, although they typically provide better predictive performance.

Embedded methods combine the characteristics of both filter and wrapper methods. They consider feature interactions similarly to wrapper methods while maintaining computational efficiency closer to that of filter methods. In embedded methods, feature subset selection is integrated directly into the training process, meaning that model learning and feature selection are performed simultaneously rather than separately.

In the literature, feature selection has been shown to significantly influence the performance of one-class classification models, as demonstrated by Hancock et al. (2025), Schneider et al. (2025), Seliya et al. (2021), and Wang and Qiao (2023). However, these studies have two major limitations. First, none of them consider all three types of feature selection methods in a unified performance comparison, leaving open the question of which method is most suitable for one-class classifiers. Second, they do not address applications in the medical domain. Moreover, the comparative effects of feature selection and instance selection, i.e. one of the key objectives of this paper, have not been systematically examined, limiting our understanding of effective data preprocessing strategies for OCC-based medical classification.

2.2 Instance Selection

Instance selection is an outlier-handling technique aimed at removing redundant or unrepresentative samples to improve model performance. When addressing class imbalance problems, this method removes a portion of samples from both the majority and minority classes to reduce noise and outliers, thereby improving classification accuracy and preventing negative effects on model learning (Garcia et al., 2012; Olvera-Lopez et al., 2010).

In general, instance selection methods can be divided into three categories: condensation, edition, and hybrid methods. Condensation methods retain samples close to the decision boundary while removing

internal samples. Edition methods remove noisy samples or those inconsistent with their neighbors to smooth the decision boundary. Hybrid methods remove both noisy and redundant samples while retaining those that are important for classification.

In related work, instance selection has been shown to play an important role in one-class classification. For example, Czarnowski (2022) combined instance selection, over-sampling, and one-class classification for class-imbalanced datasets; Li et al. (2025) integrated instance selection with different one-class classifiers for ocean observatory data; and Wang et al. (2021) combined instance selection, missing-value imputation, and one-class classification for incomplete datasets with class imbalance.

However, one major limitation of these works is that the effects of different types of instance selection algorithms on one-class classifiers for class-imbalanced medical datasets have not been fully explored. Moreover, the performance differences between feature selection and instance selection for one-class classifiers have not been examined.

2.3 One-Class Classification

In one-class classification (OCC) methods, one class (referred to as the target class or positive class) is correctly labeled in the training dataset, whereas the other class (referred to as the negative class or outlier class) consists of unlabeled data or categories that are extremely rare or even absent from the training data. For example, in automatic disease detection, it is relatively easy to collect samples representing the positive class (common diseases), while samples of the negative class (rare diseases) are often scarce. Therefore, the objective is to build a classification model using a large number of positive samples so that it can distinguish positive samples from negative ones, thereby enabling effective detection and prevention (Hayashi et al., 2026; Seliya et al., 2021).

OCC methods can generally be categorized into density estimation methods, reconstruction methods, and boundary methods. Density estimation methods employ variations of nearest-neighbor approaches to identify clusters of neighboring points. Points located far from the boundaries of surrounding clusters are considered anomalies or outliers. A representative algorithm in this category is the Local Outlier Factor (LOF) method (Breunig et al., 2000).

Boundary methods aim to determine the boundary of the target class and identify the most appropriate decision boundary. Compared with density estimation methods, boundary-based approaches typically require fewer target-class samples to define an effective boundary. Popular methods include the One-Class Support Vector Machine (OCSVM) and Support Vector Data Description (SVDD) (Alam et al., 2020).

The Isolation Forest (IF) method, which is based on the concept of decision trees, applies the Random Forest framework to outlier detection (Liu et al., 2008). Unlike distance- or density-based approaches, it isolates anomalies through random partitioning, which reduces training time and achieves linear time complexity. Because it can be trained using subsets of data while maintaining strong performance on large datasets, it is well suited for online anomaly detection.

Overall, despite the progress made in applying feature selection and instance selection to one-class classification, several limitations remain in existing studies. Previous research has generally examined these preprocessing strategies separately, without providing a unified comparison of their relative effectiveness under consistent experimental settings. Furthermore, many studies focus on only a limited range of algorithms or specific application domains, which restricts the generalizability of their findings. In particular, comprehensive investigations that jointly evaluate different feature selection and instance selection approaches for one-class classifiers in class-imbalanced medical datasets are still lacking. As a result, the comparative impact of these preprocessing techniques on OCC performance remains insufficiently understood, highlighting the need for a systematic and integrated evaluation framework.

3. Research Methodology

3.1 Data Preprocessing Procedures

This study employs four data preprocessing procedures to generate different feature representations for constructing one-class classification (OCC) models. These procedures include feature selection (FS), instance selection (IS), and their combinations in two different orders: FS followed by IS (FSIS) and IS followed by FS (ISFS). Figure 1 illustrates these procedures, which are described as follows.

- Baseline: The baseline approach utilizes the original training set (*TR*) without any preprocessing to construct OCC classifiers, including one-class support vector machine (OCSVM), isolation forest (IF), and local outlier factor (LOF) (see Section 3.2.4). The testing set (*TE*) is then used to evaluate the performance of these classifiers.
- Feature Selection (FS): The FS procedure applies feature selection algorithms, including classification and regression tree (CART), correlation-based feature selection (CFS), and genetic algorithm (GA) (see Section 3.2.2), to remove irrelevant or redundant features from *TR*. This results in three feature subsets, denoted as TR^{FS}_{CART}, TR^{FS}_{CFS}, and TR^{FS}_{GA}, which are used to train the OCC classifiers. For testing, the corresponding features are selected from *TE*, yielding TE^{FS}_{CART}, TE^{FS}_{CFS}, and TE^{FS}_{GA}.
- Instance Selection (IS): The IS procedure employs instance selection algorithms, including condensed nearest neighbor (CNN), edited nearest neighbor (ENN), and genetic algorithm (GA) (see Section 3.2.3), to remove noisy or redundant samples from TR. This produces three reduced training subsets, denoted as TR^{IS}_{CNN}, TR^{IS}_{ENN}, and TR^{IS}_{GA}, which are used to train the OCC classifiers. The testing set *TE* remains unchanged and is used for performance evaluation.
- FSIS (Feature Selection followed by Instance Selection): The FSIS procedure first applies feature selection to *TR*, followed by instance selection on the resulting feature-reduced dataset. For example, applying CART followed by CNN produces a reduced dataset denoted as $TR^{FSIS}_{CART-CNN}$.
- ISFS (Instance Selection followed by Feature Selection): The ISFS procedure performs instance selection first on *TR*, followed by feature selection on the reduced dataset. For example, applying

CNN followed by CART yields a dataset denoted as $TR^{ISFS}_{CNN-CART}$. It is noteworthy that, despite the use of identical feature selection and instance selection algorithms, variations in the order of application (FSIS vs. ISFS) may lead to differences in the resulting numbers of features and instances.

Table 1 presents computational-complexity expressions for the baseline, feature selection (FS), instance selection (IS), FSIS, and ISFS procedures. Let *TR* and *TE* denote the training and test sets; let *F_A*(*TR*) be the subset of features selected by feature-selection algorithm *A*, and let *I_B*(*TR*) be the subset of training instances retained by instance-selection algorithm *B*. Then FSIS means *I_B*(*F_A*(*TR*)), whereas ISFS means *F_A*(*I_B*(*TR*)). Feature selection determined on *TR* must be applied unchanged to *TE*. Here *n* and *d* denote the original numbers of training instances and features, and *n*″ and *d*′ the corresponding retained quantities.

Table 1
Computational complexities of different procedures.

Procedures	Big O
Baseline	$O\left(nd\right)$
FS	λ CART: $O\left(ndlogn\right)$ λ CFS: $O\left(nd^2\right)$ λ GA: $O\left(P_fG_fnd\right)$ P_f: population size; G_f: number of generations
IS	λ CNN: $O\left(n^2d\right)$ λ ENN: $O\left(n^2d\right)$ λ GA: $O\left(P_iG_in^2d\right)$ P_i: population size; G_i: number of generations
FSIS	$T_{FSIS}=T_{FS}\left(n,d\right)+T_{IS}(n,d')$ For the example of CART + CNN: $O\left(ndlogn\right)$ + $O\left(n^2d\prime\right)$
ISFS	$T_{ISFS}=T_{IS}\left(n,d\right)+T_{FS}(n\prime,d)$ For the example of CNN + CART: $O\left(n^2d\right)$ + $O\left(n\prime dlogn\prime\right)$

3.2 Experimental Setup

3.2.1 The Medical Datasets

In this paper, we focus on hard-to-classify medical datasets obtained from the KEEL dataset repository, which provides a wide range of benchmark datasets for class imbalance research. Among these datasets, 22 belong to the medical domain. Furthermore, each dataset is partitioned into 80% training and 20% testing sets using 5-fold cross-validation, ensuring fair comparisons and supporting reliable experimental conclusions.

Specifically, we define a dataset as hard-to-classify if at least two of the three OCC classifiers (i.e. OCSVM, IF, and LOF) achieve an area under the receiver operating characteristic (ROC) curve (AUC) below 0.7. Table 2 shows the AUC results of these classifiers over the 22 medical datasets.

Table 2
Baseline area under the ROC curve (AUC) for three one-class classifiers across 22 KEEL datasets.

Datasets	OCSVM	IF	LOF
Ecoli-0_vs_1	0.945	0.978	0.943
Ecoli-0-1-3-7_vs_2–6	0.827	0.812	0.830
Ecoli 1	0.818	0.860	0.885
Ecoli 2	0.725	0.748	0.823
Ecoli 3	0.748	0.755	0.599
Ecoli 4	0.881	0.874	0.895
Haberman	0.534	0.666	0.664
New-thyroid1	0.895	0.937	0.937
New-thyroid2	0.891	0.944	0.937
Pima	0.699	0.730	0.694
Wisconsin	0.994	0.995	0.984
Yeast-2_vs_8	0.774	0.663	0.782
Yeast1	0.519	0.542	0.617
Yeast3	0.592	0.697	0.816
Yeast4	0.766	0.752	0.681
Yeast5	0.906	0.908	0.704
Yeast6	0.790	0.751	0.616
Yeast-0-5-6-7-9_vs_4	0.721	0.685	0.713
Yeast-1_vs_7	0.658	0.619	0.646
Yeast-1-2-8-9_vs_7	0.636	0.624	0.620
Yeast-1-4-5-8_vs_7	0.543	0.504	0.512
Yeast-2_vs_4	0.879	0.919	0.903

Based on this criterion and Table 2, seven KEEL datasets are selected for further analysis. The basic characteristics of these datasets are summarized in Table 3, where they are ordered according to their class imbalance ratios. Notably, these datasets represent three distinct medical problems: protein localization site prediction (Yeast series), breast cancer prediction (Haberman), and diabetes prediction (Pima).

Table 3
Sample sizes, feature counts, and class-imbalance ratios of seven selected KEEL datasets.

Datasets	Imbalance ratios	No. of instances	No. of features
Pima	1.87	768	8
Yeast1	2.46	1484	8
Haberman	2.78	306	3
Yeast3	8.10	1484	8
Yeast-1_vs_7	14.30	459	7
Yeast-1-4-5-8_vs_7	22.10	693	8
Yeast-1-2-8-9_vs_7	30.57	947	8

3.2.2 Feature Selection Algorithms

Three feature selection algorithms are employed including classification and regression tree (CART), correlation-based feature selection (CFS), and genetic algorithm (GA), which belong to embedded, filter, and wrapper methods, respectively. Related parameters of implementing these algorithms are shown in Table 4.

Table 4
Parameter settings for classification and regression tree (CART), correlation-based feature selection (CFS), and genetic algorithm (GA) feature selection.

Feature selection algorithms	Parameters
CART	criterion: gini; random_state: 42
CFS	Search strategy: best first search (BFS) max_no_improve: 5
GA	n_gen: 20; pop_size: 30; cxpb: 0.6; mutpb: 0.03; indpb: 0.05; selection: tournament selection; tournsize: 3

3.2.3 Instance Selection Algorithms

Three instance selection algorithms are selected for performance comparison: CNN (Condensed Nearest Neighbor), ENN (Edited Nearest Neighbor), and GA (Genetic Algorithm), which belong to condensation, edition, and hybrid methods, respectively. These algorithms are implemented using the NoiseFiltersR software package, and their corresponding parameter settings are summarized in Table 5. It should be noted that the parameters for CNN, ENN, and GA follow the recommendations provided by their original authors, namely Tomek (1976), Wilson (1972), and Kuncheva (1995), respectively.

Table 5
Parameter settings for condensed nearest neighbor (CNN), edited nearest neighbor (ENN), and genetic algorithm (GA) instance selection.

Instance selection algorithms	Parameters
CNN	nearest neighbors, i.e. K = 3
ENN	nearest neighbors, i.e. K = 3
GA	NGEN = 50; CXPB = 0.7; MUTB = 0.2; n = 20; tournsize = 3; indpb = 0.05; weights(1.0,1.0); num_runs = 5

3.2.4 One-Class Classification Algorithms

In this paper, three OCC classifiers are developed for performance comparison, which are one-class support vector machine (OCSVM), isolation forest (IF), and local outlier factor (LOF). Table 6 shows hyperparameter to tune for these classifiers.

Table 6
Candidate hyperparameter values for one-class support vector machine (OCSVM), isolation forest (IF), and local outlier factor (LOF).

Classification algorithms	Parameters
OCSVM	Kernel: rbf; gamma: scale; nu: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5]
IF	n_estimators: [50, 100, 150, 200]; contamination: [auto, 0.1, 0.2, 0.3, 0.4, 0.5]
LOF	n_neighbors: [10, 20, 30, 40, 50]; contamination: [auto, 0.1, 0.2, 0.3, 0.4, 0.5]; novelty: True

To evaluate the performance of these OCC classifiers, the Area Under the Curve (AUC) is used, as it is a well-established and widely adopted metric in the literature on imbalanced learning. In addition, recall is assessed as a critical metric in anomaly detection, since it reflects a model's ability to correctly identify minority (anomalous) instances. For example, in cancer prediction, models with high recall demonstrate a stronger ability to correctly identify malignant cases among benign ones. Similarly, in diabetes prediction, high recall indicates the model's effectiveness in correctly detecting diabetic cases from non-diabetic individuals.

4. Experimental Results

4.1 Results on Feature Selection and Instance Selection Methods

4.1.1 The AUC Rates

Tables 7 to 9 present the AUC results for the baseline model as well as feature selection (FS) and instance selection (IS) methods applied to the OCSVM, IF, and LOF classifiers, respectively. The numbers in parentheses following the average values indicate ranks among the baseline and six preprocessing variants (1 = highest mean). Original underlining denotes best values within each dataset row.

Table 7
AUC of the OCSVM classifier: unprocessed baseline, feature selection (FS), and instance selection (IS).

Datasets	**Baseline**	**FS**			**IS**		
		CART	***CFS***	***GA***	***CNN***	***ENN***	***GA***
Haberman	0.534	0.476	0.645	0.526	0.569	0.512	0.543
Pima	0.699	0.676	0.680	0.696	0.734	0.670	0.708
Yeast1	0.519	0.540	0.546	0.514	0.586	0.453	0.529
Yeast3	0.592	0.881	0.906	0.702	0.619	0.405	0.601
Yeast-1_vs_7	0.658	0.714	0.660	0.644	0.676	0.567	0.665
Yeast-1-2-8-9_vs_7	0.636	0.692	0.638	0.659	0.640	0.582	0.638
Yeast-1-4-5-8_vs_7	0.543	0.566	0.506	0.535	0.550	0.488	0.551
Avg.	0.597 (6)	0.649 (2)	0.654 (1)	0.611 (4)	0.625 (3)	0.525 (7)	0.605 (5)

Table 8
AUC of the IF classifier: unprocessed baseline, feature selection (FS), and instance selection (IS).

Datasets	**Baseline**	**FS**			**IS**		
		CART	***CFS***	***GA***	***CNN***	***ENN***	***GA***
Haberman	0.666	0.553	0.696	0.661	0.707	0.632	0.679
Pima	0.730	0.683	0.708	0.734	0.767	0.692	0.742
Yeast1	0.542	0.558	0.576	0.549	0.629	0.481	0.558
Yeast3	0.697	0.890	0.907	0.787	0.745	0.487	0.713
Yeast-1_vs_7	0.619	0.708	0.633	0.630	0.650	0.502	0.634
Yeast-1-2-8-9_vs_7	0.624	0.669	0.650	0.632	0.605	0.517	0.606
Yeast-1-4-5-8_vs_7	0.504	0.542	0.482	0.489	0.489	0.429	0.507
Avg.	0.626 (6)	0.658 (2)	0.665 (1)	0.640 (4)	0.656 (3)	0.534 (7)	0.634 (5)

Table 9
AUC of the LOF classifier: unprocessed baseline, feature selection (FS), and instance selection (IS).

Datasets	**Baseline**	**FS**			**IS**		
		CART	***CFS***	***GA***	***CNN***	***ENN***	***GA***
Haberman	0.664	0.546	0.631	0.678	0.700	0.639	0.678
Pima	0.694	0.666	0.668	0.698	0.737	0.637	0.700
Yeast1	0.617	0.643	0.640	0.615	0.696	0.504	0.633
Yeast3	0.816	0.937	0.807	0.891	0.852	0.487	0.806
Yeast-1_vs_7	0.646	0.712	0.697	0.614	0.664	0.492	0.651
Yeast-1-2-8-9_vs_7	0.620	0.654	0.653	0.653	0.643	0.546	0.626
Yeast-1-4-5-8_vs_7	0.512	0.557	0.456	0.490	0.528	0.448	0.518
Avg.	0.653 (5)	0.674 (2)	0.650 (6)	0.663 (3)	0.689 (1)	0.536 (7)	0.659 (4)

Overall, the results indicate that applying preprocessing techniques, either FS or IS, generally improves classification performance compared to the baseline across most datasets. However, the degree of improvement varies depending on the classifier type, preprocessing strategy, and dataset characteristics.

From the average AUC values, LOF consistently achieves the highest performance among the three classifiers, followed by IF, while OCSVM generally performs the worst. This suggests that density-based methods (LOF) are more robust for the datasets considered, particularly under preprocessing conditions.

Specifically, LOF achieves its best average AUC of 0.689 using CNN (instance selection). IF achieves its best average AUC of 0.665 using CFS (feature selection), while OCSVM achieves its best average AUC of 0.654 also using CFS. These findings highlight that feature selection methods are more beneficial for OCSVM and IF, whereas instance selection is more advantageous for LOF.

Among preprocessing techniques, CNN (instance selection) yields the best overall performance, particularly when combined with LOF. CART (feature selection) ranks second and provides competitive results. CFS performs well for IF and OCSVM but less effectively for LOF. ENN and GA-based instance selection methods consistently show lower performance, indicating that aggressive instance filtering may remove informative samples.

A key observation is the impact of imbalance ratio (IR) on preprocessing effectiveness. For datasets with low imbalance ratios (Haberman, Pima, and Yeast1), CNN significantly outperforms CART. In contrast, for datasets with high imbalance ratios, CART performs better. This indicates that instance selection is more effective when class distributions are relatively balanced, while feature selection becomes more important in highly imbalanced scenarios.

Figure 2 summarizes the best-performing preprocessing methods under different class imbalance conditions. As shown in Fig. 2(a), CNN combined with LOF achieves the highest average AUC for datasets with low imbalance ratios. In contrast, Fig. 2(b) shows that CART combined with LOF achieves the highest average AUC for datasets with high imbalance ratios. These findings suggest that no single preprocessing method is universally optimal and that the choice of preprocessing strategy should consider dataset characteristics.

1. datasets with low imbalance ratios

(b) datasets with high imbalance ratios

Figure 2. Best average AUC values obtained using the baseline, feature selection (FS), and instance selection (IS) methods for (a) datasets with low imbalance ratios and (b) datasets with high imbalance ratios.

Additionally, LOF benefits more from instance selection because it relies on local density estimation. Removing noisy or redundant instances improves neighborhood structure and enhances detection performance. In contrast, feature selection helps reduce dimensionality and noise, which benefits global models such as OCSVM and IF.

4.1.2 The Recall Rates

Tables 10 to 12 present the recall results of the baseline model as well as feature selection (FS) and instance selection (IS) methods applied to the OCSVM, IF, and LOF classifiers, respectively.

Table 10
Recall of the OCSVM classifier: unprocessed baseline, feature selection (FS), and instance selection (IS).

Datasets	**Baseline**	**FS**			**IS**		
		CART	***CFS***	***GA***	***CNN***	***ENN***	***GA***
Haberman	0.532	0.507	0.753	0.557	0.579	0.446	0.593
Pima	0.768	0.728	0.720	0.757	0.929	0.683	0.809
Yeast1	0.531	0.538	0.541	0.513	0.629	0.534	0.552
Yeast3	0.637	0.982	0.969	0.828	0.717	0.454	0.668
Yeast-1_vs_7	0.733	0.667	0.667	0.667	0.733	0.667	0.733
Yeast-1-2-8-9_vs_7	0.700	0.733	0.667	0.633	0.700	0.700	0.700
Yeast-1-4-5-8_vs_7	0.533	0.667	0.467	0.533	0.600	0.533	0.600
Avg.	0.634 (6)	0.689 (2)	0.683 (3)	0.641 (5)	0.698 (1)	0.574 (7)	0.665 (4)

Table 11
Recall of the IF classifier: unprocessed baseline, feature selection (FS), and instance selection (IS).

Datasets	**Baseline**	**FS**			**IS**		
		CART	***CFS***	***GA***	***CNN***	***ENN***	***GA***
Haberman	0.816	0.691	0.765	0.829	0.902	0.729	0.865
Pima	0.840	0.750	0.780	0.836	0.970	0.765	0.877
Yeast1	0.594	0.601	0.627	0.594	0.708	0.583	0.629
Yeast3	0.858	0.976	0.975	0.902	0.938	0.674	0.896
Yeast-1_vs_7	0.633	0.767	0.700	0.667	0.700	0.600	0.700
Yeast-1-2-8-9_vs_7	0.600	0.700	0.700	0.667	0.600	0.600	0.667
Yeast-1-4-5-8_vs_7	0.500	0.567	0.400	0.500	0.567	0.533	0.567
Avg.	0.692 (6)	0.722 (3)	0.707 (5)	0.714 (4)	0.769 (1)	0.641 (7)	0.743 (2)

Table 12
Recall of the LOF classifier: unprocessed baseline, feature selection (FS), and instance selection (IS).

Datasets	**Baseline**	**FS**			**IS**		
		CART	***CFS***	***GA***	***CNN***	***ENN***	***GA***
Haberman	0.815	0.604	0.765	0.779	0.902	0.643	0.829
Pima	0.780	0.712	0.743	0.803	0.944	0.679	0.832
Yeast1	0.716	0.730	0.730	0.718	0.841	0.587	0.741
Yeast3	0.939	0.988	0.920	0.976	0.957	0.606	0.927
Yeast-1_vs_7	0.700	0.800	0.733	0.733	0.700	0.667	0.733
Yeast-1-2-8-9_vs_7	0.667	0.733	0.633	0.700	0.700	0.733	0.700
Yeast-1-4-5-8_vs_7	0.633	0.667	0.500	0.633	0.633	0.667	0.533
Avg.	0.750 (4)	0.748 (5)	0.718 (7)	0.763 (2)	0.811 (1)	0.655 (6)	0.756 (3)

Based on the average recall values, LOF achieves the highest recall performance overall, with a maximum average of 0.811 using CNN (instance selection). IF ranks second, achieving a best average recall of 0.769 also using CNN. OCSVM shows the lowest recall performance, with its best average recall of 0.698 using CNN. This consistent trend indicates that LOF is particularly effective in identifying minority class instances.

Among preprocessing techniques, CNN (instance selection) clearly provides the best performance across all three classifiers. CART (feature selection) performs competitively, especially for IF and LOF, but does not surpass CNN. CFS shows moderate performance but is less consistent. ENN and GA-based instance selection methods yield the lowest recall values, suggesting that excessive instance removal negatively impacts the detection of minority instances.

The impact of class imbalance is also evident. For datasets with low imbalance ratios (Haberman, Pima, and Yeast1), CNN significantly improves recall across all classifiers. In contrast, for datasets with high imbalance ratios, CART becomes more competitive and, in some cases, outperforms CNN. This indicates that instance selection is more effective for relatively balanced datasets, while feature selection is more beneficial for highly imbalanced scenarios.

Figure 3 compares the best average recall rates obtained using different preprocessing methods under varying class imbalance conditions. As illustrated in Fig. 3(a), CNN combined with LOF achieves the highest average recall for datasets with low imbalance ratios. In contrast, Fig. 3(b) shows that CART combined with LOF achieves the highest average recall for datasets with high imbalance ratios. These

results further suggest that preprocessing effectiveness depends on dataset characteristics, particularly the degree of class imbalance.

(a) datasets with low imbalance ratios

(b) datasets with high imbalance ratios

Figure 3. Best average recall values obtained using the baseline, feature selection (FS), and instance selection (IS) methods for (a) datasets with low imbalance ratios and (b) datasets with high imbalance ratios.

Additionally, recall is highly sensitive to instance selection because it depends on correctly identifying minority samples. CNN preserves representative and boundary instances, improving detection capability. In contrast, feature selection reduces dimensionality and noise, benefiting classifiers differently.

In summary, preprocessing significantly improves recall performance, particularly with instance selection methods. CNN combined with LOF yields the best overall results, while CART provides a strong alternative for highly imbalanced datasets. These findings highlight the importance of selecting preprocessing strategies based on dataset characteristics and evaluation goals.

4.2 Results on Different Combinations of Feature Selection and Instance Selection Methods

As LOF outperforms the other classifiers under most data preprocessing methods, only the results of FSIS and ISFS using the LOF classifier are reported due to space limitations.

4.2.1 The FSIS Method

Tables 13 and 14 present the AUC and recall results of the LOF classifier under the FSIS (Feature Selection followed by Instance Selection) method, respectively. Overall, the results indicate that combining feature selection and instance selection can further enhance anomaly detection performance compared to using either technique individually. However, the effectiveness of FSIS depends strongly on the specific combination of feature selection (FS) and instance selection (IS) methods.

Table 13
AUC of the local outlier factor (LOF) classifier with feature selection followed by instance selection (FSIS).

Datasets	CART→IS			CFS→IS			GA→IS		
	ENN	*CNN*	*GA*	*ENN*	*CNN*	*GA*	*ENN*	*CNN*	*GA*
Haberman	0.609	0.534	0.536	0.685	0.667	0.705	0.659	0.649	0.674
Pima	0.750	0.627	0.710	0.749	0.628	0.680	0.756	0.624	0.718
Yeast1	0.703	0.525	0.652	0.689	0.538	0.637	0.672	0.515	0.609
Yeast3	0.956	0.369	0.938	0.810	0.201	0.893	0.923	0.525	0.884
Yeast-1_vs_7	0.720	0.676	0.718	0.730	0.518	0.666	0.653	0.496	0.614
Yeast-1-2-8-9_vs_7	0.674	0.622	0.677	0.661	0.622	0.640	0.696	0.602	0.652
Yeast-1-4-5-8_vs_7	0.562	0.523	0.574	0.554	0.461	0.461	0.552	0.453	0.489
Avg.	0.711 (1)	0.554 (7)	0.686 (4)	0.697 (3)	0.519 (9)	0.669 (5)	0.702 (2)	0.552 (8)	0.663 (6)

Table 14
Recall of the local outlier factor (LOF) classifier with feature selection followed by instance selection (FSIS).

Datasets	CART→IS			CFS→IS			GA→IS		
	ENN	*CNN*	*GA*	*ENN*	*CNN*	*GA*	*ENN*	*CNN*	*GA*
Haberman	0.813	0.629	0.616	0.840	0.765	0.925	0.889	0.693	0.816
Pima	0.877	0.702	0.791	0.922	0.687	0.717	0.929	0.694	0.851
Yeast1	0.848	0.604	0.758	0.830	0.636	0.762	0.797	0.624	0.716
Yeast3	0.988	0.828	0.988	0.882	0.724	0.963	0.988	0.754	0.964
Yeast-1_vs_7	0.867	0.767	0.800	0.767	0.733	0.733	0.767	0.600	0.667
Yeast-1-2-8-9_vs_7	0.733	0.733	0.733	0.667	0.733	0.633	0.767	0.733	0.700
Yeast-1-4-5-8_vs_7	0.667	0.533	0.600	0.667	0.500	0.467	0.600	0.533	0.533
Avg.	0.828 (1)	0.685 (7)	0.755 (4)	0.796 (3)	0.683 (8)	0.743 (6)	0.820 (2)	0.662 (9)	0.750 (5)

Table 13 shows that CART–ENN has the highest mean AUC (0.711), followed by GA–ENN (0.702) and CFS–ENN (0.697). CFS–CNN has the lowest mean AUC (0.519). These observations depend on the particular sequence and algorithms used.

As shown in Table 14, CART–ENN achieves the highest mean recall (0.828), followed by GA–ENN (0.820) and CFS–ENN (0.796). GA–CNN records the lowest mean recall (0.662).

A key observation is that ENN-based instance selection performs better than CNN in the FSIS framework, which contrasts with the findings in Section 4.1 where CNN generally outperformed ENN. This suggests that when feature selection is applied first, the data becomes more structured and less noisy, allowing ENN to effectively refine the dataset by removing borderline or mislabeled instances without significantly harming minority class representation.

The outcomes do not establish that one feature selection method is superior in every combination: CART–ENN leads these FSIS comparisons, while the effect of feature selection depends on the subsequent instance selection method.

Within FSIS, CART–ENN provides the highest observed mean AUC and recall for LOF. However, other sequences may perform better on individual datasets; these descriptive comparisons do not establish statistical significance.

4.2.2 The ISFS Method

Tables 15 and 16 present the AUC and recall results of the LOF classifier under the ISFS (Instance Selection followed by Feature Selection) method, respectively. In contrast to FSIS, the ISFS approach applies instance selection prior to feature selection, which leads to different performance characteristics. Overall, the results show that ISFS can also improve detection performance, but its effectiveness is generally lower and less consistent compared to FSIS.

Table 15
AUC of the local outlier factor (LOF) classifier with instance selection followed by feature selection (ISFS).

Datasets	ENN→FS			CNN→FS			GA→FS		
	CART	*CFS*	*GA*	*CART*	*CFS*	*GA*	*CART*	*CFS*	*GA*
Haberman	0.635	0.705	0.700	0.483	0.621	0.634	0.600	0.626	0.678
Pima	0.754	0.762	0.741	0.639	0.617	0.638	0.703	0.701	0.698
Yeast1	0.692	0.692	0.705	0.522	0.544	0.508	0.637	0.642	0.635
Yeast3	0.953	0.814	0.899	0.610	0.678	0.617	0.944	0.887	0.875
Yeast-1_vs_7	0.730	0.724	0.647	0.619	0.555	0.557	0.588	0.547	0.622
Yeast-1-2-8-9_vs_7	0.676	0.658	0.695	0.669	0.640	0.628	0.699	0.674	0.576
Yeast-1-4-5-8_vs_7	0.577	0.478	0.515	0.575	0.493	0.460	0.542	0.558	0.517
Avg.	0.717 (1)	0.690 (3)	0.700 (2)	0.588 (8)	0.593 (7)	0.577 (9)	0.673 (4)	0.662 (5)	0.657 (6)

Table 16
Recall of the local outlier factor (LOF) classifier with instance selection followed by feature selection (ISFS).

Datasets	ENN→FS			CNN→FS			GA→FS		
	CART	*CFS*	*GA*	*CART*	*CFS*	*GA*	*CART*	*CFS*	*GA*
Haberman	0.827	0.790	0.902	0.518	0.690	0.727	0.702	0.740	0.829
Pima	0.914	0.940	0.944	0.690	0.702	0.695	0.765	0.776	0.829
Yeast1	0.809	0.792	0.830	0.604	0.657	0.620	0.736	0.746	0.755
Yeast3	0.994	0.933	0.982	0.871	0.920	0.846	0.988	0.975	0.976
Yeast-1_vs_7	0.767	0.800	0.733	0.733	0.567	0.600	0.700	0.600	0.733
Yeast-1-2-8-9_vs_7	0.733	0.633	0.800	0.733	0.667	0.700	0.767	0.667	0.600
Yeast-1-4-5-8_vs_7	0.667	0.533	0.567	0.633	0.633	0.600	0.567	0.633	0.533
Avg.	0.816 (2)	0.774 (3)	0.823 (1)	0.683 (9)	0.691 (7)	0.684 (8)	0.746 (5)	0.734 (6)	0.751 (4)

Table 15 shows that ENN–CART provides the highest average AUC (0.717), followed by ENN–GA (0.700) and ENN–CFS (0.690). The nine combinations are reported in the order shown in the table; they should not be confused with feature-selection-first sequences.

Table 16 reports the highest average recall in the third column (0.823), followed by the first (0.816) and second (0.774). The source table has conflicting first- and second-level headings; therefore, the exact feature-selection labels attached to these scores require verification against the analysis output before the associated method can be named conclusively.

A key difference from FSIS is that instance selection performed first may remove informative samples before feature selection is applied, which can negatively impact the identification of relevant features. This is especially problematic for minority class instances, as their removal reduces the ability of subsequent feature selection methods to capture meaningful patterns.

Additionally, while ENN and GA perform well as initial instance selection methods, their effectiveness depends on the preservation of sufficient representative samples. CNN, although effective in Section 4.1, shows more variable performance in ISFS, suggesting that its aggressive filtering may not be suitable when followed by feature selection.

The ISFS results are heterogeneous across combinations. The data do not support a general assertion that CART is always less effective when applied after instance selection; performance depends on the specific sequence.

In summary, the highest observed mean AUC among the listed ISFS combinations is 0.717 (ENN–CART), slightly above the highest FSIS mean AUC (0.711, CART–ENN). The highest FSIS mean recall is 0.828; Table 16 reports a highest ISFS mean recall of 0.823, subject to confirmation of its combination label.

Figure 4 summarizes the best-performing baseline, FS, IS, FSIS, and ISFS methods in terms of average AUC and recall rates using the LOF classifier. Overall, the figure provides a comprehensive comparison of individual preprocessing techniques and their combinations, highlighting the impact of both method selection and processing order on anomaly detection performance.

(a) average AUC rates

(b) average recall rates

Figure 4. Comparison of the best-performing baseline, FS, IS, FSIS, and ISFS methods using LOF in terms of (a) average AUC and (b) average recall.

Figure 4(a) compares the highest observed mean AUC values for LOF across preprocessing procedures. ENN–CART (ISFS, 0.717) and CART–ENN (FSIS, 0.711) are close in magnitude; a statistical significance claim cannot be made without a suitable test.

Figure 4(b) shows that CART–ENN (FSIS) has a mean recall of 0.828, compared with the largest value shown in Table 16 (ISFS, 0.823). Since Table 16 contains conflicting combination headings, its top-performing sequence should be confirmed before attaching a definitive method name.

5. Conclusion

This study systematically investigated the impact of preprocessing-induced feature representation changes on one-class classification performance for class-imbalanced medical datasets. Specifically, we examined the effectiveness of feature selection, instance selection, and their sequential combinations under different preprocessing orders using seven hard-to-classify medical datasets from the KEEL repository. Three representative OCC algorithms, namely OCSVM, IF, and LOF, were evaluated using AUC and recall as performance metrics.

The experimental results demonstrate that data preprocessing significantly improves OCC performance compared with baseline models across most datasets. Among the evaluated OCC methods, LOF consistently achieves the best overall performance, indicating that density-based approaches are particularly suitable for medical anomaly detection tasks under class imbalance. Regarding individual preprocessing techniques, CNN-based instance selection provides the strongest improvement overall, especially for datasets with relatively low imbalance ratios, whereas CART-based feature selection performs better for highly imbalanced datasets. These findings suggest that the effectiveness of preprocessing strategies is closely related to dataset characteristics and imbalance severity.

Combined preprocessing can outperform individual preprocessing for LOF, but its benefit depends on the algorithm pairing and order. CART–ENN (FSIS) achieves mean AUC and recall of 0.711 and 0.828, whereas ENN–CART (ISFS) achieves the highest reported mean AUC of 0.717. Accordingly, the current results do not support a universal preference for FSIS over ISFS.

Overall, this study fills an important research gap by providing one of the first comprehensive evaluations of how preprocessing affects OCC performance in medical datasets. The findings offer practical guidance for selecting preprocessing strategies when designing OCC-based diagnostic systems for imbalanced medical data.

Future research may extend this work by investigating deep learning-based feature representation techniques, exploring larger and more diverse medical datasets, and integrating adaptive preprocessing mechanisms into automated OCC frameworks to further improve robustness and generalizability.

Declarations

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The work was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 115-2410-H-182 -022 -MY2, and in part by the Chang Gung Memorial Hospital at Linkou, under Grant BMRPH13.

Author Contribution

Conceptualization: Min-Wei HuangMethodology: Wei-Chao LinSoftware: Chih-Fong TsaiValidation: Wei-Chao LinData curation: Chih-Fong TsaiProject administration: Min-Wei HuangWriting – Original Draft: Min-Wei Huang, Chih-Fong Tsai, and Wei-Chao LinWriting – Review & Editing: Min-Wei Huang, Chih-Fong Tsai, and Wei-Chao Lin

Acknowledgements

Not applicable

Availability of data and materials

The data that support the findings are publicly accessible at https://sci2s.ugr.es/keel/datasets.php

References

1. Alam, S., Sonbhadra, S.K., Agarwal, S., Nagabhushan, P.: One-class support vector classifiers: a survey. Knowl. Based Syst. **196**, 105754 (2020)
2. Araf, I., Idri, A., Chairi, I.: Cost-sensitive learning for imbalanced medical data: a review. Artif. Intell. Rev. **57**, 80 (2024)

3. Aubaidan, B.H., Kadir, R.A., Ghafoor, K.: A review of intelligent data analysis: machine learning approaches for addressing class imbalance in healthcare – challenges and perspectives. *Intelligent Data Analysis*, vol. 29, no. 3, pp. 699–719. (2025)
4. Baro, P., Borah, M.D.: GAN-based approach for data imputation and handling class imbalance using one class ensemble. Appl. Soft Comput. **182**, 113540 (2025)
5. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. *ACM SIGMOD International Conference on Management of Data*, pp. 93–104. (2000)
6. Czarnowski, I.: Weighted ensemble with one-class classification and over-sampling and instance selection (WECOI): an approach for learning from imbalanced data streams. J. Comput. Sci. **61**, 101614 (2022)
7. Dhal, P., Azad, C.: A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581. (2022)
8. Fu, Y., Liu, K., Zhu, W.: Instance selection via Voronoi neighbors for binary classification tasks. IEEE Trans. Knowl. Data Eng. **36**, 3921–3933 (2024)
9. Garcia, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 417–435 (2012)
10. Guimaraes, V., Sousa, I., Cunha, R., Magalhaes, R., Machado, A., Fernandes, V., Reis, S., Correia, M.V.: One-class classification with confound control for cognitive screening in older adults using gait, fingertapping, cognitive, and dual tasks. *Computer Methods and Programs in Biomedicine*, vol. 259, 108508. (2025)
11. Hancock, J.T., Khoshgoftaar, T.M., Liang, Q.: A problem-agnostic approach to feature selection and analysis using SHAP. J. Big Data, **12**, article 12. (2025)
12. Hayashi, T., Fujita, H.: One-class ensemble classifiers for data imbalance problems. Appl. Intell. **52**, 17073–17089 (2022)
13. Hayashi, T., Cimr, D., Fujita, H., Cimler, R.: Critical review for one-class classification: recent advances and reality behind them. WIREs Data Min. Knowl. Discov., **16**, 1, e70058. (2026)
14. Huang, M.-W., Tsai, C.-F., Wu, W.-S.: Ensemble Instance Selection for medical datasets: single-stage parallel and two-stage serial ensemble approaches. *Concurrency and Computation: Practice and Experience*, vol. 38, e70478. (2026)
15. Jian, J., Xu, L., Gong, C., Ding, S., Gong, X., Yua, X., Zheng, W., Wang, X., Zhang, Y.: One class classification-empowered radiomics for noninvasively accurate prediction of glioma isocitrate dehydrogenase mutation using multiparametric magnetic resonance imaging. Clin. Radiol. **84**, 106866 (2025)
16. Kuncheva, L.I.: Editing for the k-nearest neighbors rule by a genetic algorithm. Pattern Recognit. Lett. **16**, 809–814 (1995)
17. Leevy, J., Hancock, J., Khoshgoftaar, T.M., Zadeh, A.A.: Evaluating the impact of majority versus minority class training for one-class classification. Int. J. Artif. Intell. Tools. **34**, 4–5 (2025)

18. Li, N., Qi, Y., Xin, R., Zhu, P., Zhao, Z.: Leveraging active learning for ocean data quality assessment: reducing labeling workload and addressing severe data imbalance challenges. Int. J. Data Sci. Analytics. **20**, 4777–4798 (2025)
19. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. *IEEE International Conference on Data Mining*, pp. 413–422. (2008)
20. Marques, H.O., Swersky, L., Sander, J., Campello, R.J.G.B., Zimek, A.: On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. *Data Mining and Knowledge Discovery*, vol. 37, pp. 1473–1517. (2023)
21. Moslemi, A.: A tutorial-based survey on feature selection: recent advancements on feature selection. *Engineering Applications on Artificial Intelligence*, vol. 126, part D, 107136. (2023)
22. Olvera-Lopez, J., Carrasco-Ochoa, J.A., Martinez-Trinidad, J.F., Kittler, J.: A review of instance selection methods. Artif. Intell. Rev. **34**, 133–143 (2010)
23. Ouzar, Y., Ajmi, F., Othman, S.B., Rousseliere, C., Decaudin, B., Odou, P., Hammadi, S.: Interpretable one-class classification framework for prescription error detection using BERT embeddings and dimensionality reduction. *Computers in Biology and Medicine*, vol. 196, part B, 110775. (2025)
24. Qin, L., Wang, X., Ying, L., Jiang, Z.: A distributed evolutionary based instance selection algorithm for big data using Apache Spark. Applied Soft Computing, **159**, 111638. (2024)
25. Salmi, M., Atif, D., Oliva, D., Abraham, A., Ventura, S.: Handling imbalanced medical datasets: review of a decade of research. Artif. Intell. Rev. **57**, 273 (2024)
26. Schneider, M.P., Malegori, C., Gomes, A.D.A., Oliveri, P.: Enhancing one-class classification performance through variable selection: a review based on advanced literature search approaches. Chemometr. Intell. Lab. Syst. **265**, 105491 (2025)
27. Seliya, N., Zadeh, A.A., Khoshgoftaar, T.M.: A literature review on one-class classification and its potential applications in big data. Journal Big Data. **8**, 122 (2021)
28. Somaratne, U.V., Wong, K.W., Parry, J., Laga, H.: The use of generative adversarial networks for multi-site one-class follicular lymphoma classification. Neural Comput. Appl. **35**, 20569–20579 (2023)
29. Theng, D., Bhoyar, K.K.: Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl. Inf. Syst. **66**, 1575–1637 (2023)
30. Tomek, I.: Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772. (1976)
31. Wang, Z., Qiao, X.: Set-valued classification with out-of-distribution detection for many classes. J. Mach. Learn. Res. **24**, 1–39 (2023)
32. Wang, Z., Tsai, C.-F., Lin, W.-C.: Data cleaning issues in class imbalanced datasets: instance selection and missing values imputation for one-class classifiers. Data Technol. Appl. **55**(5), 771–787 (2021)
33. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man. Cybernetics. **SMC–2**(3), 408–421 (1972)

34. Woldaregay, A.Z., Launonen, I.K., Albers, D., Igual, J., Arsand, E., Hartvigsen, G.: A novel approach for continuous health status monitoring and automatic detection of infection incidences in people with type 1 diabetes using machine learning algorithms (part 2): a personalized digital infectious disease detection mechanism. J. Med. Internet Res. **22**, 8, e18912 (2020)
35. Yu, L., Wang, Y., Zhou, L., Wu, J., Wang, Z.: Residual neural network-assisted one-class classification algorithm for melanoma recognition with imbalanced data. Comput. Intell. **39**(6), 1004–1021 (2023)
36. Zhao, Y., Yuan, C., Liang, Y., Li, Y., Li, C., Zhao, M., Hu, J., Zhong, N., Liu, W., Liu, C.: Streamlining thoracic radiotherapy quality assurance: on-class classification for automated OAR contour assessment. Technol. Cancer Res. Treat. (2025). 10.1177/15330338251345895

Figures

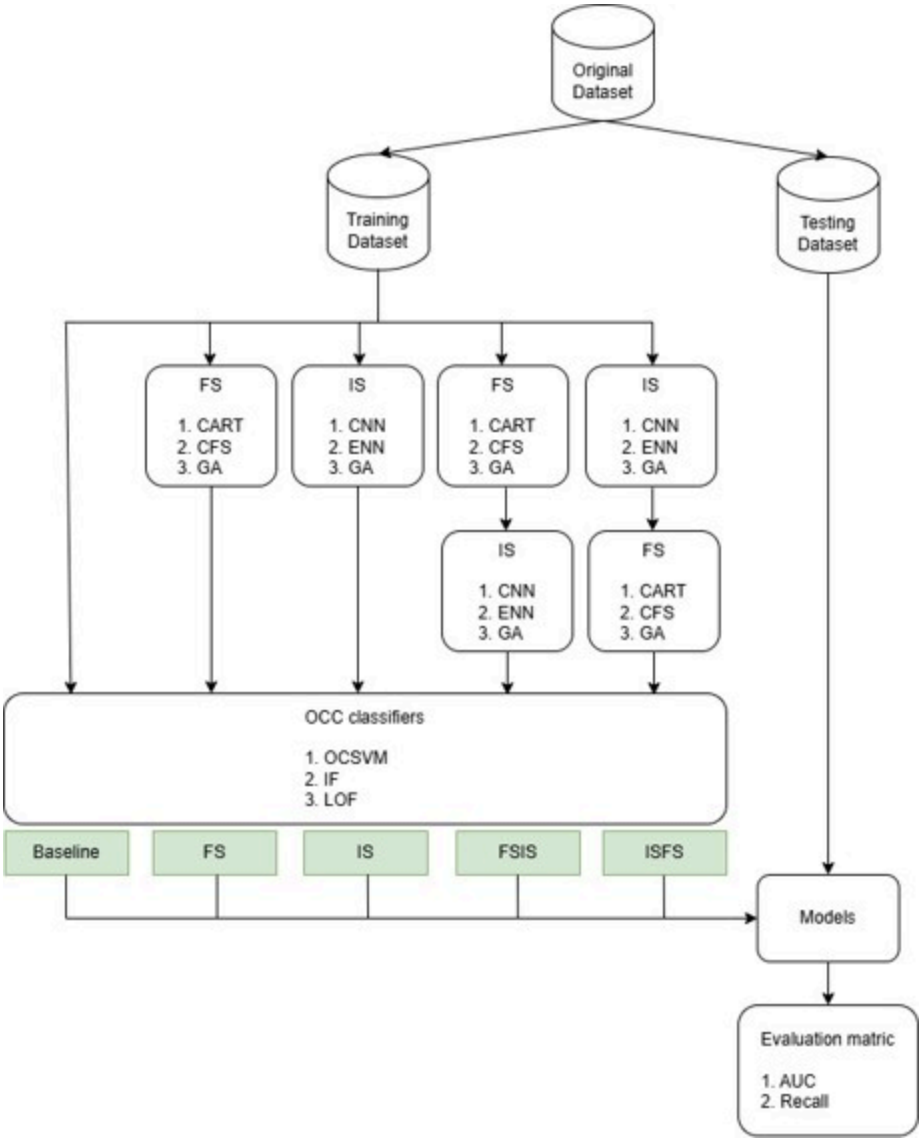
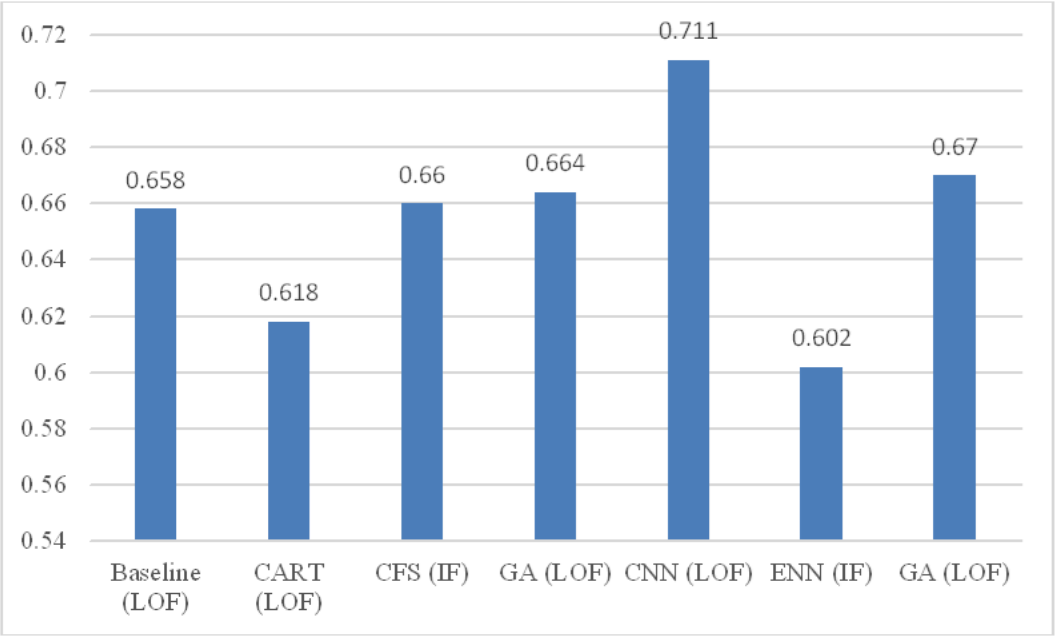
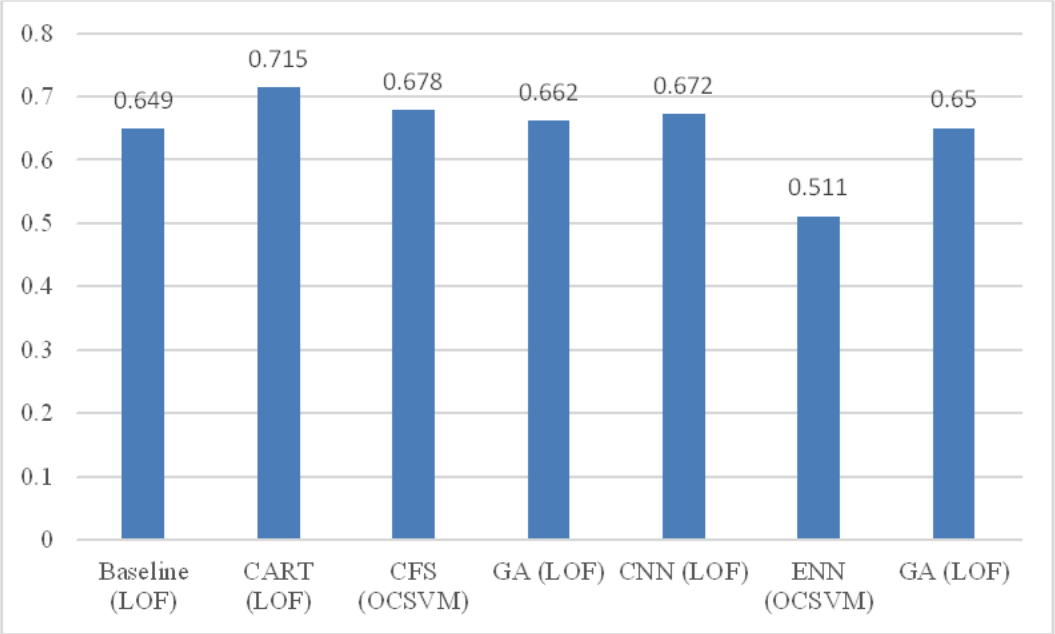


Figure 1

Data preprocessing procedures



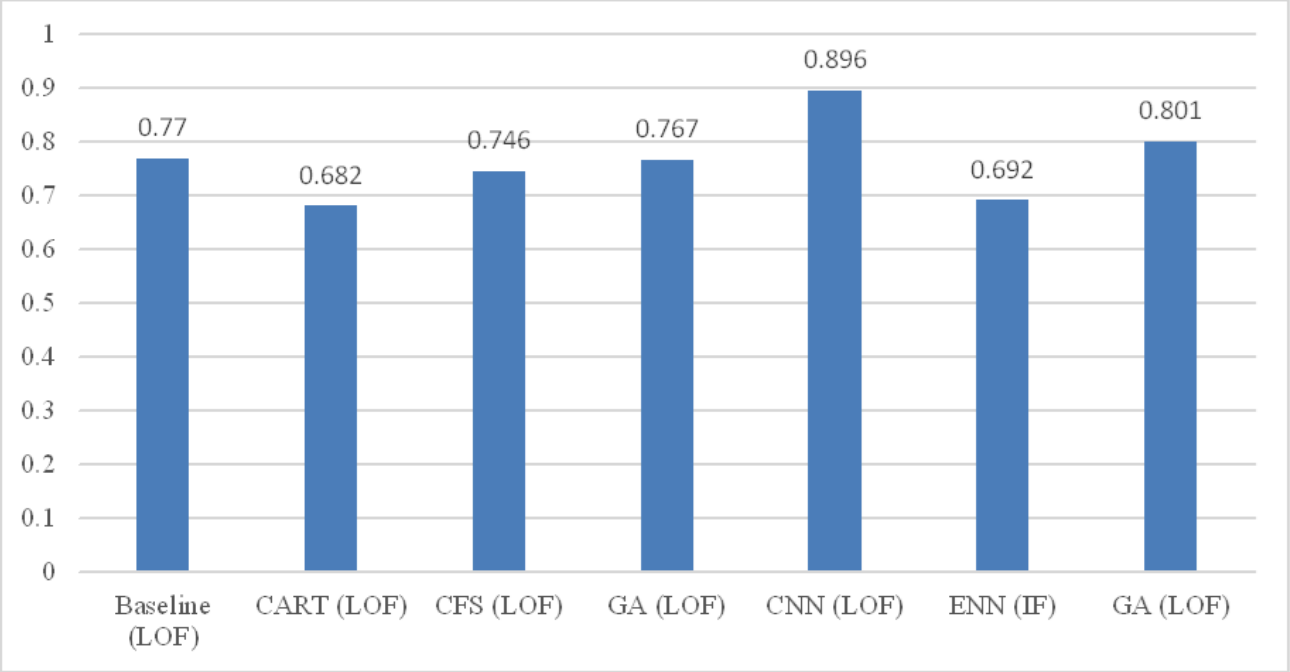
(a) datasets with low imbalance ratios



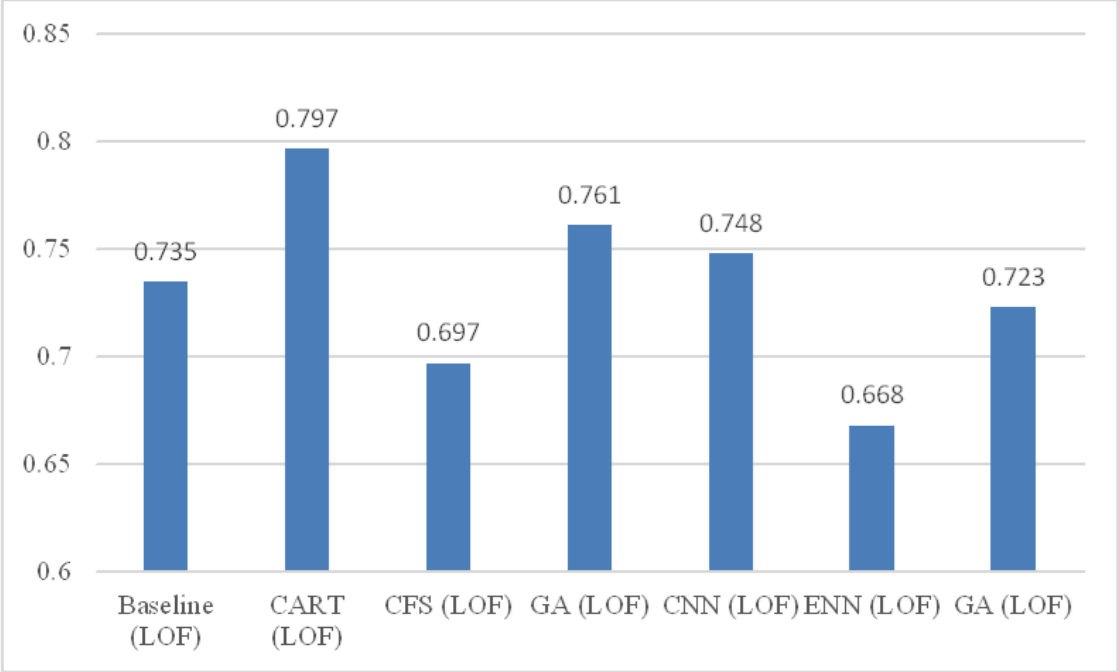
(b) datasets with high imbalance ratios

Figure 2

Best average AUC values obtained using the baseline, feature selection (FS), and instance selection (IS) methods for (a) datasets with low imbalance ratios and (b) datasets with high imbalance ratios.



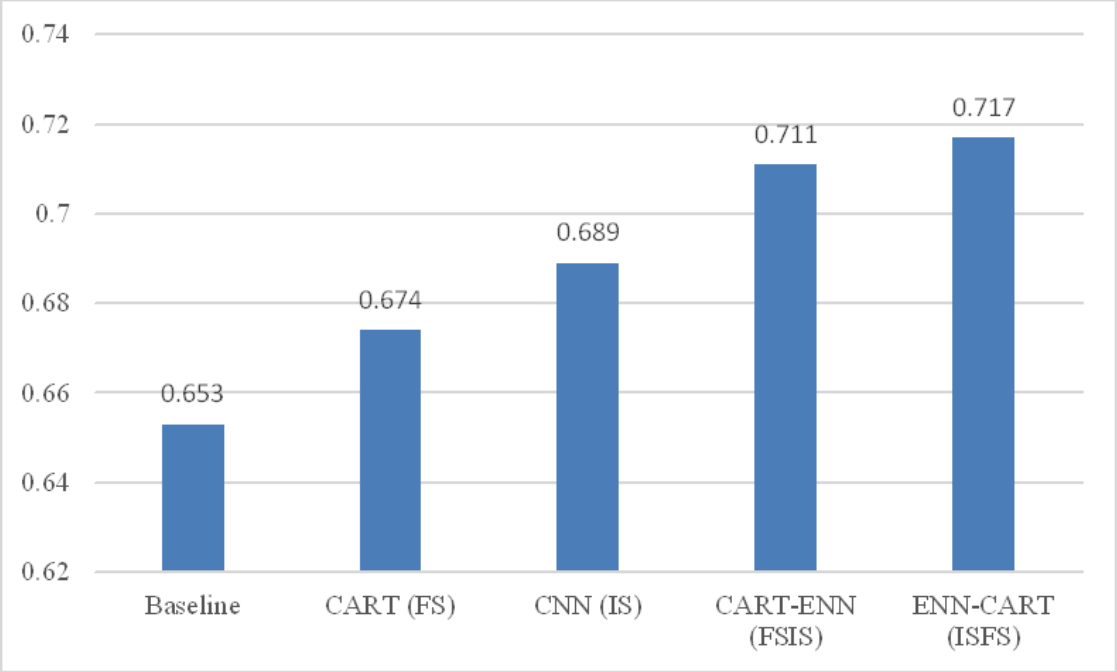
(a) datasets with low imbalance ratios



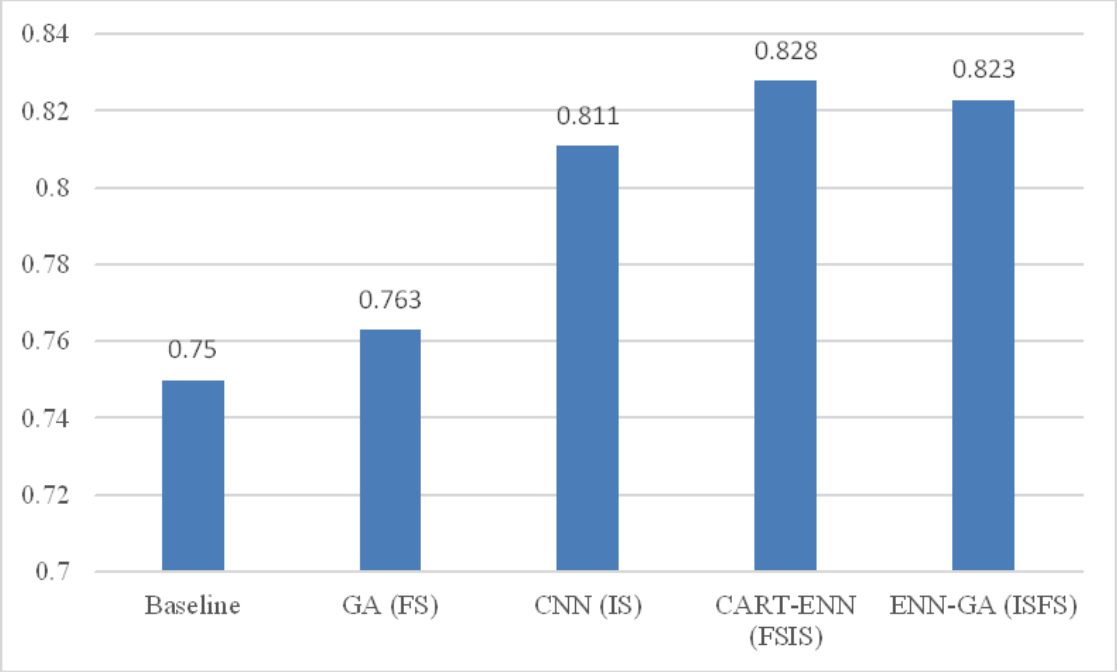
(b) datasets with high imbalance ratios

Figure 3

Best average recall values obtained using the baseline, feature selection (FS), and instance selection (IS) methods for (a) datasets with low imbalance ratios and (b) datasets with high imbalance ratios.



(a) average AUC rates



(b) average recall rates

Figure 4

Comparison of the best-performing baseline, FS, IS, FSIS, and ISFS methods using LOF in terms of (a) average AUC and (b) average recall.