Article

Not peer-reviewed version

Sensitivity-Constrained Evolutionary Feature Selection for Imbalanced Medical Classification: A Case Study on Rotator Cuff Tear Surgery Prediction

Jose María Belmonte , Fernando Jiménez * , Gracia Sánchez , Santiago Gabardo , Natalia Martínez-Catalán , Emilio Calvo , Gregorio Bernabé , José Manuel García

Posted Date: 28 October 2025

doi: 10.20944/preprints202510.2132.v1

Keywords: imbalanced classification; feature selection; evolutionary algorithms; sensitivity constraints; rotator cuff tear

 doi:10.20944/preprints202510.2132.v1

Article

Sensitivity-Constrained Evolutionary Feature Selection for Imbalanced Medical Classification: A Case Study on Rotator Cuff Tear Surgery Prediction

Jose María Belmonte [1], Fernando Jiménez [2,*], Gracia Sánchez [2], Santiago Gabardo [3,4], Natalia Martínez-Catalán [3,4], Emilio Calvo [3,4], Gregorio Bernabé [1] and José Manuel García [1]

[1] Computer Engineering Department, University of Murcia, Murcia, Spain
[2] Department of Information and Communication Engineering, University of Murcia, Murcia, Spain
[3] Shoulder and Elbow Reconstructive Surgery Unit, Department of Orthopedic Surgery and Traumatology, Hospital Universitario Fundación Jiménez Díaz, Universidad Autónoma de Madrid, Madrid, Spain
[4] IIS-Fundación Jiménez Díaz, Madrid, Spain
* Correspondence: fernan@um.es

Abstract

While most patients with degenerative rotator cuff tears respond to conservative treatment, a minority progress to surgery. To anticipate these cases under class imbalance, we propose a sensitivity-constrained evolutionary feature selection framework prioritizing surgical-class recall, benchmarked against traditional methods. Two variants are proposed: (i) a single-objective search maximizing balanced accuracy and (ii) a multi-objective search also minimizing the number of selected features. Both enforce a minimum-sensitivity constraint on the minority class to limit false negatives. The dataset includes 347 patients (66 surgical, 19%) described by 28 clinical, imaging, symptom, and functional variables. We compare against 62 widely adopted pipelines, including oversampling, undersampling, hybrid resampling, cost-sensitive classifiers, and imbalance-aware ensembles. The main metric is balanced accuracy, with surgical-class F1-score as secondary. Pairwise Wilcoxon tests with a win–loss ranking assessed statistical significance. Evolutionary models rank among the top; the multi-objective variant with a Balanced Bagging Classifier performs best, achieving mean balanced accuracy of 0.741. Selected subsets recurrently include age, tear location/severity, comorbidities, and pain/functional scores, matching clinical expectations. The constraint preserved minority-class recall without discarding or synthesizing data. Sensitivity-constrained evolutionary feature selection thus offers a data-preserving, interpretable solution for pre-surgical decision support, improving balanced performance and supporting safer triage decisions.

Keywords: imbalanced classification; feature selection; evolutionary algorithms; sensitivity constraints; rotator cuff tear

1. Introduction

In real-world clinical scenarios, datasets used for predictive modeling frequently exhibit class imbalance, particularly in medicine, where adverse outcomes such as mortality, disease recurrence, or treatment failure occur less frequently than their absence [1]. This imbalance poses significant challenges, as it often leads *machine learning* (ML) models to become biased toward the majority class, resulting in poor identification of critical but underrepresented cases [2]. Addressing this issue is essential in clinical settings, where the misclassification of minority cases can have severe consequences for patient care.

This study focuses on predicting the need for surgical intervention in patients with *rotator cuff tears* (RCTs), one of the most common causes of shoulder pain and functional limitation in the adult

 doi:10.20944/preprints202510.2132.v1

population [3–5]. RCTs are most commonly the result of degenerative processes, with prevalence increasing markedly with age. [6–8]. While many patients respond positively to conservative treatment such as physical therapy, analgesics, or corticosteroid injections, a substantial subset experience persistent pain, restricted mobility, and progressive tendon degeneration, ultimately requiring surgical repair [9]. Identifying, as early as possible, those patients who are likely to fail conservative treatment can improve clinical outcomes by preventing tear progression and reducing chronic muscle deterioration [10–13]. Therefore, developing accurate predictive models to guide treatment decisions is of considerable interest to orthopedic surgeons and rehabilitation specialists.

From a computational perspective, this problem is framed as a binary imbalanced classification task, where the minority class corresponds to patients who will eventually require surgery. In such contexts, the use of standard accuracy metrics can be misleading, as they predominantly reflect the majority class [14]. Instead, the primary metric used is *balanced accuracy* (BA) [15], which equally weights recall for both classes, providing a fairer evaluation of performance [16].

Several strategies have been proposed to address class imbalance, broadly categorized into *data-level techniques* (e.g., oversampling, undersampling, hybrid resampling) [17] and *algorithm-level techniques* (e.g., class weighting, specialized classifiers) [18]. While effective to some extent, these approaches face limitations, such as potential information loss, increased risk of overfitting, and reduced interpretability. Our alternative is the use of *feature selection* (FS) [19], which seeks to identify the most informative variables, thereby improving model generalization and offering clinicians a clearer understanding of relevant predictors. However, FS becomes challenging in the presence of imbalance, where minority class patterns may be underrepresented in the search process.

To address both imbalance and FS, we explore the combination of specialized classifiers for imbalanced data with evolutionary algorithms (EAs), which are population-based optimization techniques inspired by natural selection. EAs are particularly suitable for FS because they efficiently navigate large and complex search spaces while avoiding local optima. Two main EA paradigms are commonly applied:

1. *Constrained Single-objective evolutionary algorithms* (EA), which optimize a single performance criterion.
2. *Constrained Multi-objective evolutionary algorithms* (MOEA), which simultaneously optimize multiple conflicting objectives, producing a set of trade-off solutions known as the Pareto front.

In this work, we implement both paradigms. The EA focuses on maximizing BA, while the MOEA jointly optimizes BA and minimizes the number of selected features to balance predictive performance and model simplicity. Clinically, minimizing false negatives—patients incorrectly predicted to improve without surgery—is paramount, as these errors may delay appropriate treatment and lead to further tear progression and deterioration of shoulder function. In contrast, false positives, while undesirable, are comparatively less harmful since they primarily result in closer follow-up or additional diagnostic procedures. This asymmetry highlights the need to prioritize the accurate detection of surgical cases. To address this, our evolutionary algorithms incorporate a clinically driven constraint with a feasibility-first strategy [20–22] that enforces a minimum sensitivity threshold for the minority class. By doing so, we ensure that no model achieves high overall performance at the expense of overlooking patients who require surgery, an unacceptable trade-off in a clinical setting [23].

Additionally, we include the *F1-score* [24] as a complementary metric, given its focus on the positive (minority) class, which corresponds to patients who require surgery. Non-parametric statistical tests are applied to determine whether observed differences between models are statistically significant, providing rigorous validation of our findings.

In summary, our main contributions are as follows:

1. We frame the prediction of surgical need in RCTs as an imbalanced classification problem with clinically asymmetric misclassification costs.
2. We propose a constrained evolutionary FS framework that prioritizes sensitivity of the minority class and BA while exploring both single- and multi-objective optimization strategies.

 doi:10.20944/preprints202510.2132.v1

3. We benchmark our framework against a broad set of imbalance-handling techniques, including data-level, algorithm-level, and hybrid approaches, using multiple classifiers.
4. We perform an experimental evaluation with multiple random seeds and statistical testing to ensure the robustness of our conclusions.
5. We provide clinical and technical interpretations of the results, demonstrating that our approach achieves superior balance between performance, generalization, and interpretability compared to traditional methods.

The rest of this paper is organized as follows: Section 2 reviews the related work on imbalanced classification, evolutionary FS and ML for RCT prediction. Section 3 describes the dataset, clinical context, and methodology, including our proposed EA/MOEA frameworks. Section 4 presents the experimental setup, performance evaluation and analyzes the results. Finally, Section 5 summarizes the main findings and discusses future research directions.

2. Related Works

Recent years have seen sustained activity across three intertwined fronts: (i) image-centric pipelines for diagnosis and segmentation (MRI, ultrasound, and radiographs) driven largely by deep learning; (ii) tabular, clinic-driven models for screening and prognosis that rely on resampling, class weighting, or imbalance-aware ensembles; and (iii) FS and metaheuristic optimization, including single- and multi-objective EAs, to improve generalization and interpretability under class imbalance.

The challenge of class imbalance in medical datasets has been extensively studied in recent years due to its adverse impact on the development of reliable predictive models. Salmi *et al.* [25] provide a review of strategies used in the past decade to address this issue in medical applications, categorizing techniques into preprocessing, algorithmic, and hybrid methods, and highlighting the persistent methodological gaps in handling rare classes within diagnostic contexts. Aubaidan *et al.* [26] complement this view with a broad overview of ML techniques for intelligent data analysis in healthcare, emphasizing how class imbalance can undermine performance metrics and raise ethical and safety concerns. Along the same lines, Ahsan *et al.* [27] conducted a systematic review of ML-based heart disease prediction methods, identifying imbalanced data as a barrier to both generalization and interpretability, with many studies prioritizing accuracy over robustness and fairness. In response to these concerns, Nnamoko *et al.* [28] proposed a selective oversampling method combining SMOTE with outlier detection for improved diabetes prediction, while Vandewiele *et al.* [29] highlighted the risk of over-optimistic results when oversampling is improperly applied before data partitioning. Together, these works illustrate the pervasiveness of the imbalance problem across diverse clinical domains and the need for methodological design when developing ML pipelines for healthcare.

ML is increasingly being applied to orthopaedics, including tasks such as fracture detection, joint disease classification, and tendon injury assessment [30]. Within this field, RCT diagnosis and prognosis have received growing attention as potential beneficiaries of data-driven approaches. Shinohara *et al.* [31] proposed several ML classifiers to predict the risk of re-tear after arthroscopic repair, identifying age and imaging findings as key predictors. Similarly, Li *et al.* [32] developed an explainable ML model to identify relevant features for predicting RCTs in outpatient settings, achieving high accuracy and integrating the best model into a clinical application. Expanding beyond diagnostic screening, Alaiti *et al.* [33] used preoperative tabular data and multiple ML algorithms to predict failure to achieve clinically meaningful improvement two years after rotator cuff repair; while most models outperformed logistic regression, overall AUCs were moderate and the study did not focus on imbalance-aware learning. Complementing these task-specific efforts, Rodríguez *et al.* [34] reviewed AI for RCT management, covering imaging-based diagnosis, segmentation, radiograph interpretation, and outcome prediction, highlighting opportunities as well as persistent challenges such as small datasets and heterogeneous clinical definitions.

In the specific context of RCTs under class imbalance, Zhang *et al.* [35] tackle postoperative re-tear prediction using a cost-sensitive, graph-based approach that embeds imbalance handling directly in

 doi:10.20944/preprints202510.2132.v1

the loss. Their method combines a pairwise associative encoder to construct edges with a GCN weight network that learns node- and class-dependent weights, optimizing a weighted logits cross-entropy with logit adjustment and regularization.

Beyond RCT-specific applications, other works investigate domain-specific strategies for addressing imbalance in imaging data. Gao *et al.* [36] developed a deep learning-based one-class classification method for medical imaging, tackling the challenge of rare event detection using perturbation-based feature extraction. Although the approach targets images rather than tabular data, it underscores the general importance of tailored imbalance strategies in healthcare applications.

In the realm of metaheuristics, evolutionary and swarm-based methods have emerged as promising tools for handling class imbalance, especially when coupled with FS. Namous *et al.* [37] explore several metaheuristic-based approaches for imbalanced binary classification problems, evaluating how different fitness functions (such as ROC AUC and G-mean) influence the selection of relevant features. They show that conventional accuracy metrics often mislead the optimization process in imbalanced settings.

In the context of FS using multi-objective optimization, Chen *et al.* [38] propose a method that integrates dominance-based initialization and duplication analysis to improve both convergence and diversity. Their approach transforms feature-label and feature-feature correlations into multi-objective terms, resulting in more robust optimization performance. Rey *et al.* [39] introduce a hybrid evolutionary algorithm combining filter-based statistics and wrapper methods, tested across 44 imbalanced datasets. Their findings identify the best-suited filter techniques for evolutionary FS and show significant improvements for imbalanced data classification using Nearest Neighbour models.

Several studies have also examined advanced evolutionary paradigms for high dimensional or large-scale problems. Li *et al.* [40] propose a dual-population cooperative evolutionary algorithm for many-objective optimization problems, designed for large-scale decision variables, and validate it on both synthetic and real-world tasks. Although not strictly in the medical domain, their approach provides insights into handling complexity in multi-objective search spaces. Saadatmand *et al.* [41] extend this line of research by introducing JSEMO, a Jaccard-based many-objective FS algorithm specifically tailored for imbalanced data. Their method incorporates set-based variation operators and a double-weighted KNN classifier, significantly improving metrics like BA and G-mean in high-dimensional spaces.

More recently, Ding *et al.* [42] propose a multistage multitasking framework for FS in imbalanced datasets, combining SMOTE with swarm intelligence methods. Their approach aims to maximize the F1-score by capturing complex feature interactions and transferring knowledge across optimization tasks. Likewise, Dhinakaran *et al.* [43] present the SKR-DMKCF framework for medical data, which couples recursive feature elimination with multi-kernel classifiers in a distributed architecture. Their system reduces memory usage and improves scalability without compromising predictive performance.

Wrapper-based evolutionary algorithms, known for their superior accuracy but high computational demands, have also been addressed. Dominico *et al.* [44] design a multi-objective wrapper FS method that integrates filter-based pre-ranking with *Differential Evolution* (DE), showing performance gains with fewer features across benchmark datasets. To mitigate the computational burden of wrapper methods, Barradas-Palmeros *et al.* [45] propose mechanisms like sampling fraction strategies and evaluation caching within a permutation-based DE framework. Their results demonstrate that performance can be preserved while significantly reducing computational time.

Finally, Ghosh *et al.* [46] address the interaction between FS and class imbalance directly by proposing a binary DE algorithm that incorporates mutual information and a class-aware weighting scheme. Their model outperforms several baselines across multiple metrics including F-measure and G-mean, highlighting the value of combining relevance and fairness in FS.

 doi:10.20944/preprints202510.2132.v1

2.1. Conclusions of Related Works

Collectively, these studies underscore the growing importance of EAs and multi-objective optimization in addressing the dual challenges of FS and class imbalance, particularly in complex, high-stakes domains like healthcare. Prior works have proposed various combinations of oversampling techniques, cost-sensitive classifiers, and metaheuristics to improve fairness and robustness, with varying degrees of interpretability. In RCT-specific literature, recent contributions emphasize image-based pipelines, explainable outpatient diagnosis, or long-term outcome prediction, yet typically do not integrate imbalance-aware, constraint-driven FS as a central mechanism. By contrast, our work (i) encodes a clinical sensitivity requirement for the minority class directly into the optimization objective, ensuring that gains are not achieved at the expense of missed surgical candidates; (ii) preserves the integrity of the original dataset by avoiding artificial sample generation or instance removal, improving performance through subset-based FS coupled with the downstream classifier; and (iii) adopts an evaluation protocol attuned to imbalance, prioritizing BA and F1. These choices yield compact, interpretable models tailored to pre-surgical decision support in RCTs, complementing existing diagnostic and prognostic studies while addressing methodological gaps identified in the literature.

3. Materials and Methods

This section presents the dataset and clinical variables used in our study, followed by a description of the methods evaluated. We first introduce the baseline approaches commonly used to address class imbalance, including data-level, algorithm-level, and hybrid strategies. Next, we describe our proposed framework based on constrained EAs for FS, covering both single- and multi-objective optimization variants. Finally, we outline the evaluation protocol and statistical tests employed to compare all methods. The overall workflow of the proposed methodology is summarized in Figure 1.

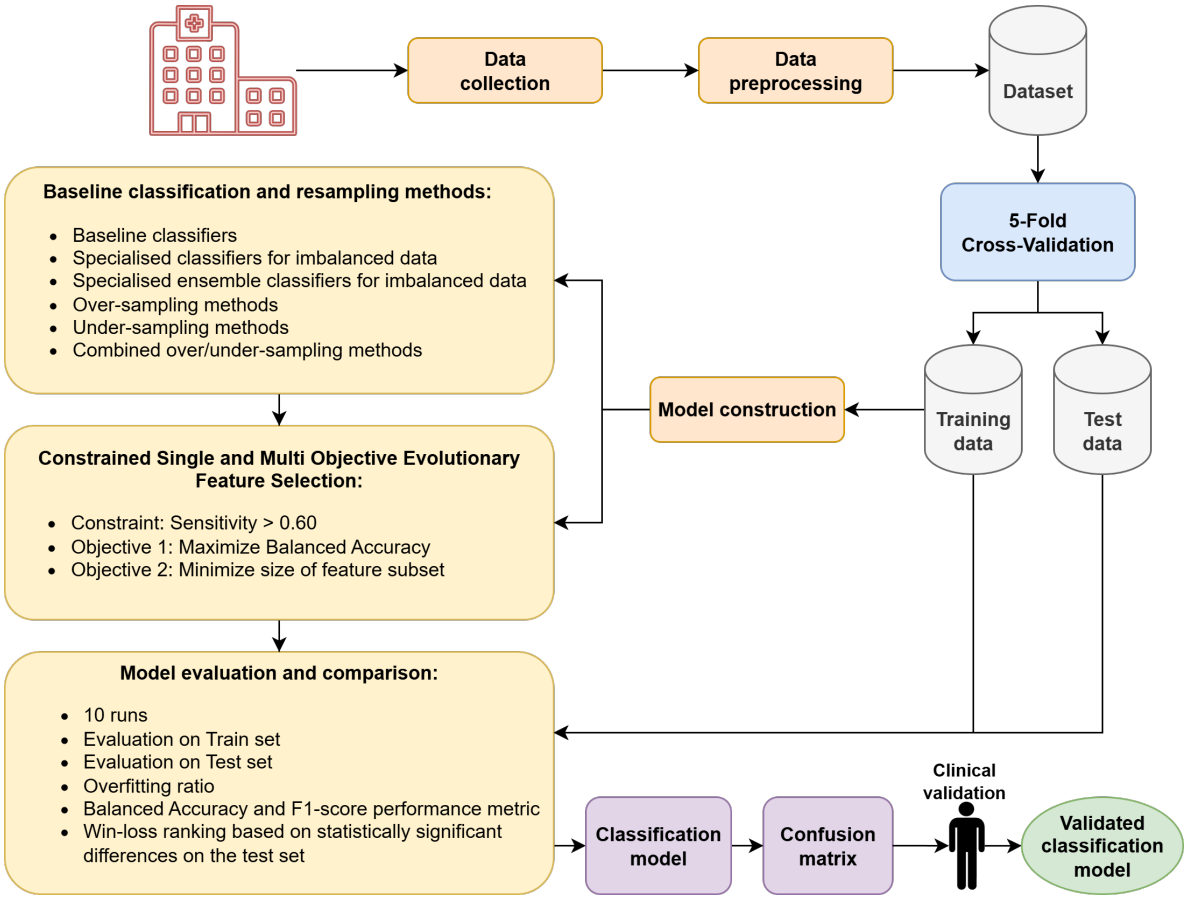

Figure 1. Flowchart of the proposed methodology.

 doi:10.20944/preprints202510.2132.v1

3.1. Dataset Description

The dataset used in this study comprises clinical, demographic, and imaging-derived variables from a prospective cohort of 347 patients diagnosed with degenerative supraspinatus RCTs. Data were collected between 2021 and 2023 under strict inclusion criteria that excluded patients with traumatic injuries, massive tears (Snyder C4), advanced fatty degeneration (Goutallier >3), or severe arthropathy (Hamada >3), ensuring a homogeneous clinical population. All patients initially underwent conservative management (e.g., physical therapy, analgesics, corticosteroid injections), with surgery considered only in those failing non-operative treatment. Of the 347 patients, 66 (19%) ultimately required surgical repair, while the remaining 281 (81%) responded successfully to conservative approaches.

The dataset includes 28 predictive features, grouped as follows:

- **Demographics and comorbidities:** age, male sex, hand dominance (affected side), manual labor occupation, smoking status, diabetes, dyslipidemia, high blood pressure, hypothyroidism, and history of rotator cuff repair on the contralateral shoulder.
- **Tear characteristics:** tear size measured in anteroposterior (mmAP) and lateral (mmLAT) planes, tear location (anterior, central, posterior thirds), complete tear larger than 20 mm, infraspinatus involvement, subscapularis tear severity (Lafosse >2), Snyder classification (C1–C3), Goutallier grade for fatty infiltration, and Hamada classification for rotator cuff arthropathy.
- **Symptoms and prior treatment:** history of corticosteroid injections, presence of night pain, and pain intensity measured via *Visual Analog Scale* (VAS).
- **Functional scores:** *Subjective Shoulder Value* (SSV) and the *American Shoulder and Elbow Surgeons* (ASES) score.

All imaging-based attributes (e.g., tear dimensions, classification scores) are assessed through *magnetic resonance imaging* (MRI), with the exception of the Hamada classification, which is derived from radiographic evaluation. Functional scores and pain assessments are recorded at the time of diagnosis based on the first MRI showing evidence of supraspinatus involvement.

No missing values were present in the dataset, as patients with incomplete records were excluded. Prior to modeling, a light preprocessing stage was applied, where categorical variables were converted into binary indicators when appropriate, and ordinal features such as tear grading or muscle atrophy classifications were encoded while preserving their intrinsic order. Feature scaling was applied exclusively to the training data, and the same transformation parameters were subsequently applied to the corresponding test set, preventing data leakage. A detailed list of input variables is provided in Table 1.

 doi:10.20944/preprints202510.2132.v1

Table 1. List of input features included in the dataset.

Attribute	Description
Age	Patient age in years
Sex (male)	Male sex (1 = yes, 0 = no)
Hand dominance	Affected side is dominant
Manual worker	Manual labor occupation
BMI	Body Mass Index
Snyder classification	Tear size grading (C1–C3)
Tear size AP	Tear size in the anteroposterior plane (mm)
Tear size LAT	Tear size in the lateral plane (mm)
Anterior location	Tear located in the anterior third
Central location	Tear located in the central third
Posterior location	Tear located in the posterior third
Complete tear >20mm	Tear larger than 20 mm
Infraspinatus involvement	Extension to infraspinatus tendon
Subscapularis involvement	Subscapularis tear (Lafosse >2)
LHBT	Long head of the biceps tendon involvement
Goutallier classification	Fatty infiltration grade
Hamada classification	Rotator Cuff arthropathy grade
Corticosteroid injections	Prior corticosteroid injections
Smoker	Smoking habits
Diabetes	Diagnosis of diabetes
Dyslipidemia	Diagnosis of dyslipidemia
High blood pressure	Diagnosis of hypertension
Hypothyroidism	Diagnosis of hypothyroidism
Contralateral side repair	Surgery on contralateral shoulder
VAS	Visual Analog Scale (pain level)
Night pain	Presence of nocturnal pain
SSV	Subjective Shoulder Value score
ASES	American Shoulder and Elbow Surgeons score
Surgery (target)	Whether the patient required surgical repair

3.2. Baseline Classifiers and Imbalance Handling Methods

To provide a broad comparison framework, we evaluated a wide range of classification algorithms and imbalance-handling strategies implemented in the *imbalanced-learn* Python package [47]. This library offers a unified interface for both data-level resampling and algorithm-level approaches, as well as specialized ensemble classifiers specifically designed for imbalanced datasets, ensuring reproducibility and standardization of experiments.

The dataset was processed using a stratified 5-fold cross-validation (CV) procedure to maintain the original class distribution in each fold. For each fold, data-level transformations were applied exclusively to the training portion to prevent information leakage, after which the resulting models were evaluated on the corresponding test set. This design allows for an unbiased assessment of each technique's ability to handle imbalance while preserving the integrity of the evaluation.

Table 2 summarizes all the evaluated techniques, organized into six main categories: (i) baseline classifiers, (ii) data-level oversampling methods, (iii) data-level undersampling methods, (iv) combined hybrid sampling techniques, (v) specialized classifiers for imbalanced datasets, and (vi) specialized ensemble classifiers for imbalanced datasets. By systematically combining resampling strategies with different classifiers, a total of 62 classification pipelines were tested. This included three baseline classifiers with six oversampling methods, eleven undersampling methods and two hybrid methods applied to each baseline classifier, two specialized imbalance-aware classifiers, and four specialized ensemble models.

 doi:10.20944/preprints202510.2132.v1

Table 2. Classification and resampling methods and their categories.

Short Name	Classifier and Parameters	Citation
Baseline classifiers		
RF	`RandomForestClassifier()`	[48]
SVC	`SVC(probability=True)`	[49]
HGB	`HistGradientBoostingClassifier()`	[50]
Oversampling methods		
ROS	`RandomOverSampler()`	[51]
SMOTE	`SMOTE()`	[52]
ADASYN	`ADASYN()`	[53]
BorderlineSMOTE	`BorderlineSMOTE()`	[54]
KMeansSMOTE	`KMeansSMOTE()`	[55]
SVMSMOTE	`SVMSMOTE()`	[56]
Undersampling methods		
CC	`ClusterCentroids()`	[57]
CNN	`CondensedNearestNeighbour()`	[58]
ENN	`EditedNearestNeighbours()`	[59]
RENN	`RepeatedEditedNearestNeighbours()`	[60]
AllKNN	`AllKNN()`	[60]
IHT	`InstanceHardnessThreshold()`	[61]
NearMiss	`NearMiss()`	[62]
NCR	`NeighbourhoodCleaningRule()`	[63]
OSS	`OneSidedSelection()`	[64]
RUS	`RandomUnderSampler()`	[65]
Tomek	`TomekLinks()`	[66]
Combined over/under-sampling methods		
SMOTEENN	`SMOTEENN()`	[67]
SMOTETomek	`SMOTETomek()`	[68]
Specialized classifiers for imbalanced data		
SVC-Balanced	`SVC(probability=True, class_weight="balanced")`	[49]
HGB-Balanced	`HistGradientBoostingClassifier(class_weight="balanced")`	[50]
Specialized ensemble classifiers for imbalanced data		
BBC	`BalancedBaggingClassifier()`	[69]
BRF	`BalancedRandomForestClassifier()`	[70]
BRF-Balanced	`BalancedRandomForestClassifier(class_weight="balanced")`	[70]

The three baseline classifiers used were *Random Forest* (RF), *Support Vector Classifier* (SVC), and *Histogram-based Gradient Boosting* (HGB). These represent three commonly used modeling paradigms: tree-based bagging ensembles, margin-based classifiers, and boosting methods, respectively. RF constructs multiple decision trees using bootstrap sampling and aggregates predictions via majority voting, providing strong performance and robustness. SVC finds an optimal separating hyperplane in a high-dimensional space and is particularly effective for complex decision boundaries. HGB leverages gradient boosting with histogram-based binning, enabling faster training and scalability on larger datasets.

To address class imbalance with these baseline classifiers, we applied a diverse set of resampling strategies. Oversampling techniques generate synthetic or replicated samples of the minority class, whereas undersampling methods reduce the number of majority class instances. Hybrid approaches, such as SMOTEENN and SMOTETomek, combine oversampling (e.g., SMOTE) with data cleaning procedures (e.g., *Edited Nearest Neighbours* or *Tomek Links*) to simultaneously increase minority representation and remove noisy samples. Each resampling method was evaluated in conjunction with the three baseline classifiers, providing a view of their combined performance. In addition to the classifiers' standard versions, these classifiers can incorporate internal mechanisms to mitigate imbalance using the `class_weight` parameter. The option `balanced` automatically assigns weights inversely proportional to class frequencies.

 doi:10.20944/preprints202510.2132.v1

Beyond individual classifiers, we included ensemble approaches specifically tailored for imbalanced learning. For example, the *Balanced Bagging Classifier* (BBC) enforces balanced class distributions within each bootstrap sample, while the *Balanced Random Forest Classifier* (BRF) is a variant of RF where each tree is trained on a balanced bootstrap subset of the data, combining the benefits of bagging and undersampling. These specialized ensembles are widely recognized for their ability to improve minority class detection without discarding large portions of the dataset.

All experiments were conducted in Python 3.10.0. Data processing and modeling relied on widely used open-source libraries, NumPy 2.1.3, Pandas 2.2.3, and Matplotlib 3.10.0, for numerical computation, data manipulation, and visualization. For ML modeling we used scikit-learn 1.6.1 and imbalanced-learn 0.13.0, and evolutionary optimization was implemented with `pymoo` 0.6.1.3.

Experiments were executed on the GACOP cluster (CPU compute node *Mendel*) with the following configuration: motherboard Supermicro X10DRL-i (SYS-6018R-MTR), dual Intel Xeon E5-2698 v4 processors, 128 GiB DDR4 2400 MHz memory (4×32 GiB DIMMs), Intel I210 Gigabit Ethernet, 1 TB HDD (ST1000NM0033-9ZM), and 80 GB Intel SSD (SSDSC2BB08). Multiple evolutionary runs and CV folds were launched in parallel across CPU cores to amortize wall-clock time.

Wall-clock times for optimization and evaluation were recorded using the standard `time` module, and per-generation/iteration progress was monitored with `tqdm` progress bars.

3.3. Constrained Evolutionary Algorithms for Feature Selection

To address the problem of imbalanced classification while reducing model complexity, we propose two wrapper-based evolutionary approaches for FS with specialized classifiers for imbalanced data: a single-objective EA and a MOEA. Both algorithms use the BA metric as the primary evaluation criterion and enforce a sensitivity-based constraint to ensure clinical reliability.

The two methods share common evolutionary components: they employ binary vectors to represent subsets of selected features, use a population size of 50 individuals, and run for 1000 generations. Each individual represents a candidate subset of features, encoded as a binary string where a bit value of 1 indicates inclusion of the corresponding feature. Evaluation is conducted using 5-fold stratified CV to estimate the model's BA on the training set.

To determine the optimal balance between exploration and exploitation, we performed a grid search over the crossover and mutation probabilities. A 10×10 search grid with three repetitions was applied, where the crossover probability was explored in the interval $(0.0, 1.0]$ and the mutation probability in the interval $(0.0, 0.3]$, with a total of 5000 evaluations. Based on this procedure, the selected crossover probability was set to 0.7, while the mutation probability was set to 0.1.

Crossover operations differ between the two approaches: the EA uses uniform crossover, whereas the MOEA applies two-point crossover. Mutation is performed using the bit-flip operator [71].

Selection in the EA is performed using ranking-based selection with an exponential ranking function, which provides adjustable selection pressure through tunable parameters, while the MOEA handles selection based on non-dominated sorting and crowding distance. Both algorithms enforce a recall-based constraint to ensure clinically acceptable sensitivity for the minority class. Infeasible solutions, those not satisfying the constraint, are penalized.

3.4. Optimization Objective and Sensitivity Constraint

Let $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$ be a binary vector representing the selection of n features, where $x_j = 1$ indicates that feature j is included in the model. Each individual $\mathbf{x}$ is evaluated using 5-fold stratified CV on the training data to ensure robustness and reduce variance.

The primary optimization objective in both the EA and MOEA is to maximize the BA. This metric is particularly well-suited for imbalanced classification problems because it equally considers the performance of both classes, mitigating the bias toward the majority class. BA is defined as:

$$\text{BA} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

 doi:10.20944/preprints202510.2132.v1

where TP (true positives) refers to correctly predicted surgical cases, FN (false negatives) to surgical cases incorrectly predicted as non-surgical, TN (true negatives) to correctly predicted non-surgical cases, and FP (false positives) to non-surgical cases incorrectly predicted as requiring surgery.

During preliminary experiments, we evaluated alternative metrics such as *specificity* (recall of the majority class) and *sensitivity* (recall of the minority class). However, these measures in isolation proved inadequate for guiding the optimization process.

In highly imbalanced datasets, a model could achieve perfect sensitivity by simply predicting every instance as belonging to the minority class, or perfect specificity by predicting exclusively the majority class. Although these extreme cases would yield high values for one metric, they would perform poorly overall and would not represent clinically useful solutions.

By combining both sensitivity and specificity, BA addresses this issue, penalizing models that overfit to a single class and rewarding those that achieve balanced performance. This makes BA a more robust measure of classifier quality, particularly when both classes have clinical importance.

Sensitivity constraint

Consider the following example with two hypothetical models:

- **Model A:** BA = 0.775. Specificity = 0.95. Sensitivity = 0.60.
- **Model B:** BA = 0.775. Specificity = 0.80. Sensitivity = 0.75.

Although both models achieve identical BA scores, Model B provides higher recall for the minority class (patients requiring surgery), resulting in fewer false negatives. In clinical practice, failing to identify patients who need surgery may delay treatment and worsen prognosis. Conversely, false positives, while undesirable, mainly lead to additional follow-up or diagnostic testing, which are comparatively less harmful.

This asymmetry highlights the need to explicitly control sensitivity to ensure that solutions are not only balanced but also clinically meaningful.

To guarantee a minimum level of detection for the minority class, we impose a sensitivity constraint during the optimization process. Only solutions that meet this minimum recall threshold are considered feasible. This ensures that the evolutionary search does not converge toward models that, while achieving high BA, neglect to correctly identify surgical cases.

Formally, the constraint is expressed as:

$$\text{Sensitivity} \geq 0.6$$

where 0.6 was chosen as a conservative lower bound based on clinical considerations and empirical evaluation.

The feasibility function used during optimization is defined as:

$$g(\mathbf{x}) = 0.6 - \text{Sensitivity}(\mathbf{x}) \leq 0$$

A feasibility-first strategy is adopted to handle this constraint, meaning that the selection mechanism prioritizes feasible solutions and compares fitness values only among them. Infeasible solutions are penalized but remain in the population to encourage exploration near the boundary of feasibility.

In addition to BA, we also monitor the *F1-score* during evaluation. The F1-score is the harmonic mean of precision and recall for the minority class and is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

 doi:10.20944/preprints202510.2132.v1

The F1-score provides a complementary perspective, as it emphasizes the minority class while also accounting for false positives. In the context of predicting surgical need, a high F1-score reflects both a low rate of missed surgical cases (high recall) and a controlled rate of unnecessary interventions (high precision). While BA serves as the optimization objective, the F1-score is used in the analysis phase to better understand the trade-offs between sensitivity and precision across different models.

To assess whether the observed performance differences between models were statistically meaningful, we perform pairwise comparisons using the *Wilcoxon signed-rank test* [72]. This non-parametric test is particularly appropriate for evaluating two related samples when no assumption of normality can be made regarding their performance differences. From the test results, we derived a win–loss ranking by counting the number of statistically significant pairwise victories (wins) and defeats (losses) for each model when compared against the others. This ranking provides a clear summary of the comparative performance of all approaches.

Single-objective evolutionary algorithm

The EA minimizes the negative BA:

$$\min F(\mathbf{x}) = -\text{BA}(\mathbf{x})$$

subject to the feasibility constraint defined above. The fitness transformation is:

$$F'(\mathbf{x}) = \begin{cases} F(\mathbf{x}) & \text{if } g(\mathbf{x}) \leq 0 \\ f_{\max} + g(\mathbf{x}) & \text{otherwise} \end{cases}$$

This structure encourages infeasible solutions to evolve toward feasibility, while driving feasible ones toward optimality. Selection is based on exponential ranking and elitism is applied to retain top individuals across generations.

Multi-objective evolutionary algorithm

The proposed MOEA is implemented using the *Non-dominated Sorting Genetic Algorithm II* (NSGA-II) scheme [73] from the `pymoo` framework [74], and is designed to optimize two conflicting objectives:

$$f_1(\mathbf{x}) = C(\mathbf{x}) \quad \text{(number of selected features)}$$
$$f_2(\mathbf{x}) = -\text{BA}(\mathbf{x}) \quad \text{(negative BA)}$$

subject to:

$$g(\mathbf{x}) = 0.6 - \text{Sensitivity}(\mathbf{x}) \leq 0$$

Each individual is evaluated to obtain a pair of objective values and a constraint violation score. Only individuals satisfying the recall threshold are considered feasible and eligible for Pareto ranking. The final output is a Pareto front of feasible, non-dominated solutions, enabling trade-offs between model interpretability and predictive performance.

4. Experimental Results and Analysis

This section presents an experimental evaluation of the proposed evolutionary feature selection frameworks, comparing our proposal against traditional data-level and model-level imbalance handling strategies. All methods were trained and evaluated following the setup described in Section 3, using the same dataset, partitions, and evaluation metrics to ensure a fair comparison.

Our analysis begins with a statistical comparison of models using non-parametric tests to determine whether observed performance differences are statistically significant, with a win–loss ranking to provide an overall performance hierarchy across all evaluated techniques. Following this, we examine the generalization capabilities of the top-performing models through the analysis of the

 doi:10.20944/preprints202510.2132.v1

overfitting ratio, comparing training and test performance. Finally, we interpret the clinical and technical implications of these results, highlighting the strengths and limitations of each approach.

4.1. Statistical Test Results

Table 3 presents the win–loss ranking for all evaluated methods, based on statistically significant differences observed on the test set.

Table 3. Win–loss ranking for the BA metric based on statistically significant differences on the test set.

Method	Wins	Losses	Wins–Losses
MOEA_BBC	67	0	67
MOEA_SVC_Bal	63	0	63
EA_SVC_Balanced	60	0	60
RUS_RF	57	0	57
IHT_RF	54	0	54
IHT_SVC	52	1	51
RUS_HGB	50	1	49
BBC	48	1	47
ENN_SVC	48	3	45
IHT_HGB	47	2	45
RENN_HGB	47	4	43
BRF	46	3	43
MOEA_BRF	44	1	43
MOEA_HGB_Bal	46	3	43
SMOTEENN_RF	46	4	42
RENN_RF	44	3	41
EA_BBC	43	2	41
EA_BRF	44	4	40
AllKNN_RF	44	6	38
CC_SVC	44	6	38
SMOTEENN_HGB	43	5	38
RUS_SVC	42	7	35
AllKNN_SVC	41	11	30
BRF_Balanced	38	8	30
SMOTEENN_SVC	34	6	28
AllKNN_HGB	33	8	25
ENN_HGB	33	14	19
RENN_SVC	37	19	18
NCR_HGB	30	13	17
ENN_RF	32	22	10
CNN_RF	24	23	1
KMeansSMOTE_SVC	23	23	0
HGB_Balanced	23	23	0
EA_HGB_Balanced	23	25	-2
Tomek_HGB	21	28	-7
SVC_Balanced	17	24	-7
SMOTE_SVC	15	24	-9
SMOTETomek_SVC	15	24	-9
ADASYN_SVC	15	25	-10
EA_BRF_Balanced	16	27	-11
KMeansSMOTE_HGB	17	30	-13
OSS_HGB	17	30	-13
NCR_RF	15	29	-14
NCR_SVC	16	31	-15
CNN_HGB	15	30	-15
SVMSMOTE_SVC	14	30	-16
ROS_HGB	14	31	-17
ROS_SVC	12	30	-18
BorderSMOTE_SVC	10	34	-24
SVMSMOTE_HGB	8	32	-24
BorderSMOTE_HGB	9	34	-25
SMOTE_HGB	8	35	-27
SMOTETomek_HGB	8	35	-27
ROS_RF	8	36	-28
CC_RF	6	36	-30
ADASYN_HGB	8	38	-30
MOEA_BRF_Bal	5	35	-30
NearMiss_RF	6	43	-37
CC_HGB	5	44	-39
ADASYN_RF	7	47	-40
BorderSMOTE_RF	8	49	-41
SMOTE_RF	6	48	-42
SMOTETomek_RF	5	48	-43
KMeansSMOTE_RF	5	50	-45
CNN_SVC	5	56	-51
SVMSMOTE_RF	5	57	-52
NearMiss_HGB	2	60	-58
OSS_RF	2	66	-64
Tomek_RF	2	66	-64
NearMiss_SVC	2	66	-64
OSS_SVC	0	70	-70
Tomek_SVC	0	70	-70

 doi:10.20944/preprints202510.2132.v1

These results demonstrate the superiority of our proposed evolutionary approaches. The MOEA combined with BBC achieves the highest overall score, with 69 wins and no losses. Similarly, MOEA and single-objective EA with SVC using balanced class weighting occupy the next top positions, highlighting the effectiveness of incorporating FS within an evolutionary optimization framework to address severe class imbalance.

In contrast, traditional resampling methods combined with standard classifiers tend to perform worse, appearing mostly in the middle or lower parts of the ranking. Undersampling methods, such as RUS or IHT, can occasionally achieve competitive performance and even rank within the top ten models. However, these approaches achieve class balance by removing a substantial number of majority class samples, which can severely reduce the amount of available training data. In real-world medical datasets, where data collection is often expensive, time-consuming, and subject to strict privacy regulations, this reduction in sample size can be particularly problematic. Eliminating potentially informative instances may lead to a loss of clinically relevant patterns, ultimately limiting the model's ability to generalize to future patient populations.

Oversampling techniques, such as SMOTE or Borderline-SMOTE, take the opposite approach by generating synthetic minority class samples to increase their representation. While this avoids the problem of discarding real patient data, it introduces artificially generated instances that may not fully capture the complex variability of real-world medical cases. As shown in Table 3, these methods tend to perform poorly, ranking near the bottom. This suggests that simply augmenting the minority class with synthetic data is insufficient for handling the complexity of this prediction task, particularly when the relationships between variables are highly non-linear and clinically nuanced.

Our evolutionary FS framework offers a distinct advantage over both resampling paradigms. By directly optimizing the selection of the most informative features, our approach achieves superior performance without altering the original dataset. Unlike undersampling, it does not discard valuable clinical information, and unlike oversampling, it does not introduce potentially unreliable synthetic data. This makes it especially suitable for medical applications, where maintaining the integrity of the dataset is crucial and where the interpretability of selected features can provide additional insights for clinicians.

To further examine the generalization capabilities of the models, we analyzed the overfitting ratio (OR), which measures the relative difference between training and test performance. Lower values indicate a smaller gap between training and test BA, suggesting better generalization and reduced risk of overfitting.

For this analysis, we selected three representative models from each category:

- Our three best-performing evolutionary approaches (EA/MOEA).
- Three strong models based on traditional undersampling techniques.
- Two specialized classifiers designed for imbalanced datasets, without FS.

Table 4 summarizes the mean and standard deviation of the BA for both training and test sets, along with the resulting OR for each model. The evolutionary methods exhibit a favorable balance, achieving competitive test performance with moderate OR values, indicating that they are not overfitting to the training data. In contrast, some specialized ensemble methods, such as BRF, display very high training accuracy but substantially lower test accuracy, leading to the highest OR values. This behavior highlights a tendency to overfit, which can limit their practical utility in real-world clinical settings.

 doi:10.20944/preprints202510.2132.v1

Table 4. Mean BA and OR for selected models. Lower OR indicates better generalization.

Model	Mean Train	Std Train	Mean Test	Std Test	OR
MOEA_BBC	0.892960	0.021218	**0.688800**	0.042791	0.228633
MOEA_SVC_Balanced	0.848060	0.022677	0.682680	0.057565	0.195010
EA_SVC_Balanced	0.865080	0.022198	0.673480	0.065626	0.221482
RUS_RF	0.871306	0.014731	0.671224	0.056455	0.229635
IHT_RF	0.773309	0.018944	0.665537	0.045253	0.139364
IHT_SVC	0.769602	0.024038	0.664025	0.042154	**0.137184**
BBC	0.926979	0.015359	0.662062	0.060425	0.285785
BRF	**0.977181**	0.007470	0.651039	0.057505	0.333758

The table shows that evolutionary models achieve a balance between high test performance and controlled overfitting, confirming their suitability for complex imbalanced classification tasks in clinical applications. On the other hand, models like BRF and BBC, although strong on training data, exhibit signs of overfitting, which can hinder their reliability when applied to unseen patient cases.

In addition to the analysis of the BA metric, we also evaluate the performance of all models using the *F1-score*, a metric particularly relevant in our clinical context since it focuses on the minority class, which represents patients requiring surgery. This metric is useful to assess how well the algorithms identify true surgical cases without excessively misclassifying non-surgical patients.

Table 5 presents the win–loss ranking for the F1-score. Similar to the BA analysis, each comparison between models is performed on the test set results across multiple runs, and statistically significant differences are recorded as wins or losses.

The F1-score results reinforce the conclusions drawn from the BA analysis. The top positions are consistently occupied by our proposed evolutionary approaches. These findings confirm that incorporating FS within an evolutionary optimization framework is highly effective for improving the identification of surgical cases.

4.2. Computational Cost and Execution Times

Evolutionary algorithms are inherently computationally intensive due to their iterative nature and the need to evaluate a population of candidate solutions across multiple generations. In our case, each individual corresponds to a subset of features, which must be trained and validated using 5-fold stratified CV.

To keep wall-clock time manageable, we executed runs in parallel on a multi-core node of our HPC cluster; wall times were recorded with the `time` module and per-generation progress monitored with `tqdm`. Table 6 reports times for each classifier for a single execution under the single-objective EA (top) and the MOEA (bottom).

In the full experimental protocol, in order to perform statistical tests, we executed 10 runs per fold and 5 folds, i.e., 50 executions per classifier. Aggregating across classifiers, the end-to-end wall-clock time to complete all experiments on our cluster was under 67 hours. These figures illustrate the expected trade-off: while evolutionary optimization incurs a higher computational cost than conventional baselines, it produces compact subsets and consistent gains in BA and F1 in a high-stakes clinical setting, justifying the one-time training investment prior to deployment.

 doi:10.20944/preprints202510.2132.v1

Table 5. Win–loss ranking for the F1-score metric based on statistically significant differences on the test set.

Method	Wins	Losses	Wins–Losses
MOEA_BBC	67	0	67
MOEA_SVC_Bal	66	0	66
EA_SVC_Balanced	58	0	58
RUS_RF	56	1	55
BBC	53	0	53
ENN_SVC	48	0	48
IHT_RF	47	2	45
BRF	47	2	45
RUS_HGB	46	2	44
MOEA_HGB_Bal	47	3	44
IHT_SVC	44	2	42
EA_BBC	44	2	42
EA_BRF	44	2	42
MOEA_BRF	44	2	42
AllKNN_RF	44	3	41
RENN_RF	41	4	37
AllKNN_SVC	42	5	37
SMOTEENN_RF	41	5	36
RENN_HGB	41	6	35
IHT_HGB	40	5	35
CC_SVC	40	7	33
BRF_Balanced	37	4	33
SMOTEENN_SVC	35	5	30
AllKNN_HGB	35	5	30
SMOTEENN_HGB	37	7	30
RUS_SVC	37	9	28
ENN_HGB	37	9	28
NCR_HGB	27	10	17
RENN_SVC	34	20	14
ENN_RF	26	14	12
KMeansSMOTE_SVC	17	15	2
CNN_RF	21	21	0
SVC_Balanced	19	19	0
HGB_Balanced	19	19	0
ADASYN_SVC	16	21	-5
EA_HGB_Balanced	21	26	-5
SMOTE_SVC	17	24	-7
SMOTETomek_SVC	17	24	-7
EA_BRF_Balanced	18	27	-9
CNN_HGB	18	28	-10
ROS_SVC	14	26	-12
Tomek_HGB	16	28	-12
CC_RF	14	27	-13
CC_HGB	14	29	-15
KMeansSMOTE_HGB	15	31	-16
OSS_HGB	13	29	-16
NCR_RF	12	30	-18
SVMSMOTE_SVC	12	30	-18
MOEA_BRF_Bal	10	28	-18
ROS_HGB	12	32	-20
NearMiss_RF	9	30	-21
NCR_SVC	11	32	-21
BorderSMOTE_SVC	11	33	-22
SVMSMOTE_HGB	7	37	-30
SMOTE_HGB	7	38	-31
SMOTETomek_HGB	7	38	-31
BorderSMOTE_HGB	7	39	-32
NearMiss_SVC	7	41	-34
ADASYN_HGB	7	44	-37
NearMiss_HGB	6	44	-38
ROS_RF	5	48	-43
ADASYN_RF	5	51	-46
BorderSMOTE_RF	5	51	-46
SMOTE_RF	4	52	-48
SMOTETomek_RF	4	52	-48
KMeansSMOTE_RF	4	59	-55
CNN_SVC	4	60	-56
SVMSMOTE_RF	4	63	-59
OSS_RF	2	68	-66
Tomek_RF	2	68	-66
OSS_SVC	0	70	-70
Tomek_SVC	0	70	-70

Table 6. Execution times by classifier for a single execution.

SVC	HGB	BRF	BRF-Balanced	BBC
EA Execution Time (hours : minutes : seconds)				
00:48:21	01:18:21	00:51:01	00:54:06	01:15:43
MOEA Execution Time (hours : minutes : seconds)				
00:38:02	01:13:26	00:51:42	00:53:04	01:05:33

 doi:10.20944/preprints202510.2132.v1

4.3. Subset-Based FS and Clinical Relevance

A central advantage of our pipeline is the explicit search over *subsets* of variables rather than relying solely on attribute *rankings*. Ranking-based tools (e.g., permutation importance, SHAP) are valuable for interpretation, but they typically assess variables in isolation or under a fixed model and can miss interactions, conditional effects, and redundancies that emerge only when features are considered jointly. By treating the predictor as a wrapper and optimizing the variable set with respect to the downstream classifier's empirical behavior, subset-based FS can (i) capture synergistic patterns among moderately informative variables, (ii) remove redundant or collinear attributes, and (iii) align the selected subset with the task objective and the sensitivity constraint used during training. This is particularly relevant in medical datasets that are often imbalanced, may contain relatively few instances, and include many attributes drawn from heterogeneous measurements (clinical scales, imaging, laboratory analyses).

4.4. Analysis of the Top Performing Models and Results Interpretation

Across the win–loss analyses based on BA and the complementary results for F1, the best overall configuration is the MOEA paired with BBC. Close contenders are the balanced SVC with both EA and MOEA, reinforcing that embedding FS within an evolutionary wrapper that optimizes BA under a clinically motivated sensitivity constraint yields consistent advantages in imbalanced settings. These results are achieved without altering the original data distribution (no over/under-sampling at the data level), which is desirable in medical cohorts where sample sizes are modest and the minority class is scarce. Averaged across the five folds, our MOEA BBC attains a final BA of 0.741, obtained by averaging the best execution across the five folds.

Comparative results highlight the limitations of pure undersampling methods (e.g., RUS, NearMiss): although they can occasionally perform well, they discard a large fraction of majority-class instances, which is problematic in real-world medical datasets where the total number of patients is limited and the minority class is small. Conversely, over-sampling approaches (e.g., SMOTE variants) avoid data loss but may inject synthetic borderline or noisy samples that do not translate into sustained gains in BA or F1 in our task. By avoiding any manipulation of the data distribution and instead optimizing a clinically guided objective, the MOEA/EA + FS strategy attains top-tier performance while preserving dataset integrity.

Importantly, MOEA consistently outperforms EA across classifiers. Since the second objective in MOEA is to reduce the number of attributes, these results suggest that our clinical dataset contains redundant or weakly relevant variables. MOEA's explicit drive toward compact subsets likely removes redundancy beyond what EA achieves, improving generalization while maintaining the sensitivity constraint.

To illustrate the learned operating point, Figure 2 reports the aggregated confusion matrix for MOEA BBC over the 5-fold CV. The model correctly identifies most Surgery cases (minority class), consistent with the sensitivity constraint. As expected, this comes with a moderate increase in false positives; in our clinical context this trade-off is acceptable, as it primarily leads to closer follow-up or additional diagnostics rather than missed surgical candidates.

The features most often selected across folds and runs are consistent with clinical expectations: Age; Snyder classification; Tear size; Tear location (anterior third); Infraspinatus involvement; comorbidities (Diabetes, Dyslipidemia, High blood pressure, Hypothyroidism); Night pain; and functional scores (SSV, ASES).

 doi:10.20944/preprints202510.2132.v1

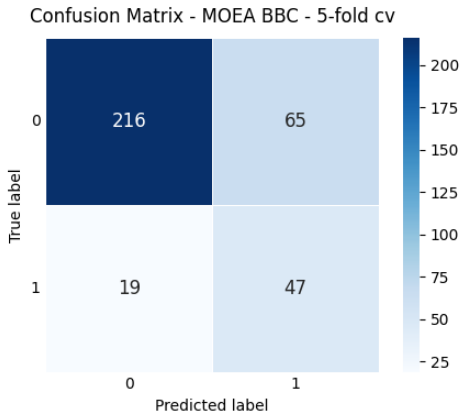

Figure 2. Aggregated confusion matrix for MOEA_BBC over 5-fold CV.

5. Conclusions and Future Work

This work presents a novel FS framework based on EAs with clinically driven constraints, aimed at addressing the challenges of imbalanced classification in healthcare applications. Our case study focused on predicting the need for surgical intervention in patients with RCTs, where early and accurate identification of individuals likely to require surgery is essential. Detecting these patients in advance allows clinicians to initiate timely interventions, prevent further tear progression, and ultimately improve patient outcomes while optimizing healthcare resource allocation.

The results demonstrate that the proposed methodology consistently outperforms traditional approaches such as data-level methods (oversampling, undersampling, and hybrid techniques) and algorithm-level strategies (class weighting and specialized classifiers). By integrating FS into the optimization process, our framework generates more interpretable models.

A key strength of our approach is that it operates directly on the original dataset without the need to modify its distribution through artificial data generation or by discarding potentially useful cases. Unlike undersampling, which removes data and may be problematic when working with small real-world datasets, or oversampling, which introduces synthetic information, our proposal preserves data integrity while focusing on identifying the most informative predictors for accurate classification. Furthermore, the incorporation of a sensitivity constraint ensures that the minority class (patients who require surgery) is not neglected, addressing the asymmetric clinical cost of misclassification.

Despite these advantages, evolutionary algorithms inherently involve a higher computational cost, as they require the evaluation of many candidate solutions across multiple generations. However, this cost is primarily incurred during the model optimization stage. Once the best solution is identified, the resulting model can be deployed efficiently for real-time predictions in clinical practice.

It is also important to acknowledge the limitations of this work. Although the dataset includes a diverse set of clinical variables, its overall size is relatively limited, underscoring the need for external and multicentric validation to ensure that the proposed models generalize well across different populations and healthcare settings.

Future directions include incorporating explainable machine learning techniques such as SHAP to enhance model transparency and support clinical decision-making. The combination of constrained optimization and explainability could lead to more trustworthy and clinically actionable AI systems. Additionally, exploring domain-specific or multiple clinical constraints (e.g., age- or risk-based thresholds) could further tailor the optimization process to practical clinical scenarios. As a complementary line, we intend to investigate strategies to accelerate evolutionary optimization, such as parallel evaluation, surrogate-assisted fitness approximation, and multi-fidelity/early-stopping schemes, so as to reduce runtime while preserving the quality of the selected feature subsets.

In conclusion, our work shows that combining FS with clinically informed constraints within an evolutionary optimization framework is a promising strategy for tackling imbalanced classification problems in medicine. This approach not only improves predictive performance but also produces

 doi:10.20944/preprints202510.2132.v1

interpretable and clinically relevant models, providing a strong foundation for future decision support systems in healthcare.

Author Contributions: Conceptualization, J.M.B., F.J., G.S., S.G., N.M.-C., E.C., G.B. and J.M.G.; methodology, J.M.B., F.J. and J.M.G.; software, J.M.B., F.J. and G.S.; validation, F.J., S.G., N.M.-C. and E.C.; formal analysis, F.J., G.S., N.M.-C. and E.C.; investigation, J.M.B., F.J., G.S., G.B. and J.M.G.; resources, S.G., N.M.-C. and E.C.; data curation, J.M.B.; writing—original draft preparation, J.M.B.; writing—review and editing, F.J., G.S., S.G., N.M.-C., E.C., G.B. and J.M.G.; supervision, F.J., G.S., S.G., N.M.-C., E.C., G.B. and J.M.G.; project administration, F.J., G.B. and J.M.G.; funding acquisition, G.B. and J.M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially funded by Grant TED2021-129221B-I00 funded by MCIN/AEI/10.13039/501100011033, and by the "European Union NextGenerationEU/PRTR".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

ADASYN	Adaptive Synthetic Sampling
AllKNN	All K-Nearest Neighbours
ASES	American Shoulder and Elbow Surgeons
BA	Balanced Accuracy
BBC	Balanced Bagging Classifier
BRF	Balanced Random Forest
CC	Cluster Centroids
CNN	Condensed Nearest Neighbour
CV	Cross-Validation
DE	Differential Evolution
EA	Evolutionary Algorithm
ENN	Edited Nearest Neighbours
FS	Feature Selection
HGB	Histogram-based Gradient Boosting
IHT	Instance Hardness Threshold
ML	Machine Learning
MOEA	Multi-Objective Evolutionary Algorithm
MRI	Magnetic Resonance Imaging
NCR	Neighbourhood Cleaning Rule
NSGA	Non-dominated Sorting Genetic Algorithm
OSS	One-Sided Selection
OR	Overfitting Ratio
RCT	Rotator Cuff Tear
RENN	Repeated Edited Nearest Neighbours
RF	Random Forest
ROS	Random Oversampling
RUS	Random Undersampling
SMOTE	Synthetic Minority Oversampling Technique
SMOTEENN	SMOTE + Edited Nearest Neighbour
SVC	Support Vector Classifier
SSV	Subjective Shoulder Value
VAS	Visual Analog Scale

 doi:10.20944/preprints202510.2132.v1

References

1. Banik, D.; Bhattacharjee, D. Mitigating data imbalance issues in medical image analysis. In *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention*; IGI Global, 2022; pp. 1215–1238. https://doi.org/10.4018/978-1-7998-7371-6.ch004.
2. Belarouci, S.; Chikh, M.A. Medical imbalanced data classification. *Advances in Science, Technology and Engineering Systems Journal* **2017**, *2*, 116–124. https://doi.org/10.25046/aj020316.
3. Nakajima, D.; Yamamoto, A.; Kobayashi, T.; Osawa, T.; Shitara, H.; Ichinose, T.; Takasawa, E.; Takagishi, K. The effects of rotator cuff tears, including shoulders without pain, on activities of daily living in the general population. *Journal of Orthopaedic Science* **2012**, *17*, 136–140. https://doi.org/10.1007/s00776-011-0186-4.
4. Keener, J.D.; Aleem, A.W.; Chamberlain, A.M.; Sefko, J.; Steger-May, K. Factors associated with choice for surgery in newly symptomatic degenerative rotator cuff tears: a prospective cohort evaluation. *Journal of Shoulder and Elbow Surgery* **2020**, *29*, 12–19. https://doi.org/10.1016/j.jse.2019.08.005.
5. Brindisino, F.; Salomon, M.; Giagio, S.; Pastore, C.; Innocenti, T. Rotator cuff repair vs. nonoperative treatment: a systematic review with meta-analysis. *Journal of Shoulder and Elbow Surgery* **2021**, *30*, 2648–2659. https://doi.org/10.1016/j.jse.2021.04.040.
6. Jung, W.; Lee, S.; Hoon Kim, S. The natural course of and risk factors for tear progression in conservatively treated full-thickness rotator cuff tears. *Journal of Shoulder and Elbow Surgery* **2020**, *29*, 1168–1176. https://doi.org/10.1016/j.jse.2019.10.027.
7. Fitzpatrick, L.; Atinga, A.; White, L.; Henry, P.; Probyn, L. Rotator Cuff Injury and Repair. *Seminars in Musculoskeletal Radiology* **2022**, *26*, 585–596. https://doi.org/10.1055/s-0042-1756167.
8. Moran, T.E.; Werner, B.C. Surgery and Rotator Cuff Disease: A Review of the Natural History, Indications, and Outcomes of Nonoperative and Operative Treatment of Rotator Cuff Tears. *Clinics in Sports Medicine* **2023**, *42*, 1–24. https://doi.org/10.1016/j.csm.2022.08.001.
9. Page, M.J.; Green, S.; McBain, B.; Surace, S.J.; Deitch, J.; Lyttle, N.; Mrocki, M.A.; Buchbinder, R. Manual therapy and exercise for rotator cuff disease. *Cochrane Database of Systematic Reviews* **2016**. https://doi.org/10.1002/14651858.CD012224.
10. Björnsson, H.C.; Norlin, R.; Johansson, K.; Adolfsson, L.E. The influence of age, delay of repair, and tendon involvement in acute rotator cuff tears. *Acta Orthopaedica* **2011**, *82*, 187–192. https://doi.org/10.3109/17453674.2011.566144.
11. Azzam, M.G.; Dugas, J.R.; Andrews, J.R.; Goldstein, S.R.; Emblom, B.A.; CainJr, E.L. Rotator Cuff Repair in Adolescent Athletes. *The American Journal of Sports Medicine* **2018**, *46*, 1084–1090. https://doi.org/10.1177/0363546517752919.
12. Schemitsch, C.; Chahal, J.; Vicente, M.; Nowak, L.; Flurin, P.H.; Lambers Heerspink, F.; Henry, P.; Nauth, A. Surgical repair versus conservative treatment and subacromial decompression for the treatment of rotator cuff tears. *The Bone & Joint Journal* **2019**, *101-B*, 1100–1106. https://doi.org/10.1302/0301-620X.101B9.BJJ-2018-1591.R1.
13. Prasetia, R.; Handoko, H.K.; Rosa, W.Y.; Ismiarto, A.F.; Petrasama.; Utoyo, G.A. Primary traumatic shoulder dislocation associated with rotator cuff tear in the elderly. *International Journal of Surgery Case Reports* **2022**, *95*, 107200. https://doi.org/10.1016/j.ijscr.2022.107200.
14. Thölke, P.; Mantilla-Ramos, Y.J.; Abdelhedi, H.; Maschke, C.; Dehgan, A.; Harel, Y.; Kemtur, A.; Mekki Berrada, L.; Sahraoui, M.; Young, T.; et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage* **2023**, *277*, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253.
15. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, 2010, pp. 3121–3124. https://doi.org/10.1109/ICPR.2010.764.
16. Chawla, N.V., Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O.; Rokach, L., Eds.; Springer US: Boston, MA, 2005; pp. 853–867. https://doi.org/10.1007/0-387-25465-X_40.
17. Mamilla, M.Y.; Al-Haddad, R.; Chowdhury, S. Resampling Imbalanced Healthcare Data for Predictive Modelling. *International Journal of Advanced Computer Science and Applications* **2025**, *16*. https://doi.org/10.14569/IJACSA.2025.0160204.
18. Yin, H.L.; Leong, T.Y. A Model Driven Approach to Imbalanced Data Sampling in Medical Decision Making. *Studies in health technology and informatics* **2010**, *160*, 856–60. https://doi.org/10.3233/978-1-60750-588-4-856.

 doi:10.20944/preprints202510.2132.v1

19. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Feature selection for high-dimensional data. *Progress in Artificial Intelligence* **2016**, *5*, 65–75. https://doi.org/10.1007/s13748-015-0080-y.
20. Jiménez, F.; Verdegay, J.L. Evolutionary techniques for constrained optimization problems. In Proceedings of the 7th European Congress on Intelligent Techniques and Soft Computing (EUFIT'99), Aachen, Germany, 1999, pp. 504–512.
21. Jiménez, F.; Verdegay, J.L.; Gómez-Skarmeta, A.F. Evolutionary techniques for constrained multiobjective optimization problems. In *Workshop on multi-criterion optimization using evolutionary methods GECCO-1999*; 1999.
22. Coello Coello, C.A. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering* **2002**, *191*, 1245–1287. https://doi.org/10.1016/S0045-7825(01)00323-1.
23. Drosou, K.; Georgiou, S.; Koukouvinos, C.; Stylianou, S. Support Vector Machines Classification on Class Imbalanced Data: A Case Study with Real Medical Data. *Journal of Data Science* **2022**, *12*, 727–754. https://doi.org/10.6339/JDS.201410_12(4).0009.
24. Powers, D.M.W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2020, [arXiv:cs.LG/2010.16061]. https://doi.org/10.48550/arXiv.2010.16061.
25. Salmi, M.; Atif, D.; Oliva, D.; Abraham, A.; Ventura, S. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review* **2024**, *57*, 273. https://doi.org/10.1007/s10462-024-10884-2.
26. Aubaidan, B.H.; Kadir, R.A.; Lajb, M.T.; Anwar, M.; Qureshi, K.N.; Taha, B.A.; Ghafoor, K. A review of intelligent data analysis: Machine learning approaches for addressing class imbalance in healthcare - challenges and perspectives. *Intelligent Data Analysis* **2025**, *29*, 699–719. https://doi.org/10.1177/1088467X241305509.
27. Ahsan, M.M.; Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine* **2022**, *128*, 102289. https://doi.org/10.1016/j.artmed.2022.102289.
28. Nnamoko, N.; Korkontzelos, I. Efficient treatment of outliers and class imbalance for diabetes prediction. *Artificial Intelligence in Medicine* **2020**, *104*, 101815. https://doi.org/10.1016/j.artmed.2020.101815.
29. Vandewiele, G.; Dehaene, I.; Kovács, G.; Sterckx, L.; Janssens, O.; Ongenae, F.; De Backere, F.; De Turck, F.; Roelens, K.; Decruyenaere, J.; et al. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine* **2021**, *111*, 101987. https://doi.org/10.1016/j.artmed.2020.101987.
30. Alzubaidi, L.; AL-Dulaimi, K.; Salhi, A.; Alammar, Z.; Fadhel, M.A.; Albahri, A.; Alamoodi, A.; Albahri, O.; Hasan, A.F.; Bai, J.; et al. Comprehensive review of deep learning in orthopaedics: Applications, challenges, trustworthiness, and fusion. *Artificial Intelligence in Medicine* **2024**, *155*, 102935. https://doi.org/10.1016/j.artmed.2024.102935.
31. Shinohara, I.; Mifune, Y.; Inui, A.; Nishimoto, H.; Yoshikawa, T.; Kato, T.; Furukawa, T.; Tanaka, S.; Kusunose, M.; Hoshino, Y.; et al. Re-tear after arthroscopic rotator cuff tear surgery: risk analysis using machine learning. *Journal of Shoulder and Elbow Surgery* **2024**, *33*, 815–822. https://doi.org/10.1016/j.jse.2023.07.017.
32. Li, C.; Alike, Y.; Hou, J.; Long, Y.; Zheng, Z.; Meng, K.; Yang, R. Machine learning model successfully identifies important clinical features for predicting outpatients with rotator cuff tears. *Knee Surgery, Sports Traumatology, Arthroscopy* **2023**, *31*, 2615–2623. https://doi.org/10.1007/s00167-022-07298-4.
33. Alaiti, R.K.; Vallio, C.S.; Assunção, J.H.; de Andrade e Silva, F.B.; Gracitelli, M.E.C.; Neto, A.A.F.; Malavolta, E.A. Using Machine Learning to Predict Nonachievement of Clinically Significant Outcomes After Rotator Cuff Repair. *Orthopaedic Journal of Sports Medicine* **2023**, *11*, 23259671231206180. https://doi.org/10.1177/23259671231206180.
34. Rodriguez, H.C.; Rust, B.; Hansen, P.Y.; Maffulli, N.; Gupta, M.; Potty, A.G.; Gupta, A. Artificial Intelligence and Machine Learning in Rotator Cuff Tears. *Sports Medicine and Arthroscopy Review* **2023**, *31*. https://doi.org/10.1097/JSA.0000000000000371.
35. Zhang, Z.; Zhang, Z.; Peng, Z.; Dong, Y. Predicting Postoperative Re-Tear of Arthroscopic Rotator Cuff Repair Using Artificial Intelligence on Imbalanced Data. *IEEE Access* **2025**, *13*, 24487–24497. https://doi.org/10.1109/ACCESS.2025.3538595.
36. Gao, L.; Zhang, L.; Liu, C.; Wu, S. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artificial Intelligence in Medicine* **2020**, *108*, 101935. https://doi.org/10.1016/j.artmed.2020.101935.

 doi:10.20944/preprints202510.2132.v1

37. Namous, F.; Faris, H.; Heidari, A.A.; Khalafat, M.; Alkhawaldeh, R.S.; Ghatasheh, N., Evolutionary and Swarm-Based Feature Selection for Imbalanced Data Classification. In *Evolutionary Machine Learning Techniques: Algorithms and Applications*; Mirjalili, S.; Faris, H.; Aljarah, I., Eds.; Springer Singapore: Singapore, 2020; pp. 231–250. https://doi.org/10.1007/978-981-32-9990-0_11.
38. Chen, C.; Yao, X.; Gong, D.; Tu, H. A multi-objective evolutionary algorithm for feature selection incorporating dominance-based initialization and duplication analysis. *Swarm and Evolutionary Computation* **2025**, *95*, 101914. https://doi.org/10.1016/j.swevo.2025.101914.
39. Rey, C.C.T.; García, V.S.; Villuendas-Rey, Y. Evolutionary feature selection for imbalanced data. In Proceedings of the 2023 Mexican International Conference on Computer Science (ENC), 2023, pp. 1–7. https://doi.org/10.1109/ENC60556.2023.10508674.
40. Li, J.; Xu, S.; Zheng, J.; Jiang, G.; Ding, W. Research on Multi-Objective Evolutionary Algorithms Based on Large-Scale Decision Variable Analysis. *Applied Sciences* **2024**, *14*. https://doi.org/10.3390/app142210309.
41. Saadatmand, H.; Akbarzadeh-T, M.R. Many-Objective Jaccard-Based Evolutionary Feature Selection for High-Dimensional Imbalanced Data Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 8820–8835. https://doi.org/10.1109/TPAMI.2024.3416196.
42. Ding, W.; Yao, H.; Huang, J.; Hou, T.; Geng, Y. Evolutionary multistage multitasking method for feature selection in imbalanced data. *Swarm and Evolutionary Computation* **2025**, *92*, 101821. https://doi.org/10.1016/j.swevo.2024.101821.
43. Dhinakaran, D.; Srinivasan, L.; Edwin Raja, S.; Valarmathi, K.; Gomathy Nayagam, M. Synergistic feature selection and distributed classification framework for high-dimensional medical data analysis. *MethodsX* **2025**, *14*, 103219. https://doi.org/10.1016/j.mex.2025.103219.
44. Dominico, G.; Bernardes, J.S.; Dorneles, L.L.; Dorn, M. Multi-Objective Wrapper Differential Evolution with Guided Initial Population for Feature Selection. In Proceedings of the 2023 IEEE Congress on Evolutionary Computation (CEC), 2023, pp. 1–8. https://doi.org/10.1109/CEC53210.2023.10254085.
45. Barradas-Palmeros, J.A.; Mezura-Montes, E.; Rivera-López, R.; Acosta-Mesa, H.G. Computational Cost Reduction in Wrapper Approaches for Feature Selection: A Case of Study Using Permutational-Based Differential Evolution. In Proceedings of the 2024 IEEE Congress on Evolutionary Computation (CEC), 2024, pp. 1–8. https://doi.org/10.1109/CEC60901.2024.10611859.
46. Ghosh, A.; Xue, B.; Zhang, M. Binary Differential Evolution based Feature Selection Method with Mutual Information for Imbalanced Classification Problems. In Proceedings of the 2021 IEEE Congress on Evolutionary Computation (CEC), 2021, pp. 794–801. https://doi.org/10.1109/CEC45853.2021.9504882.
47. Lemaitre, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, 2016. https://doi.org/10.48550/arXiv.1609.06570.
48. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
49. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*. https://doi.org/10.1145/1961189.1961199.
50. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017; NIPS'17, p. 3149–3157.
51. Kamalov, F.; Leung, H.H.; Cherukuri, A.K. Keep it simple: random oversampling for imbalanced data. In Proceedings of the 2023 Advances in Science and Engineering Technology International Conferences (ASET), 2023, pp. 1–4. https://doi.org/10.1109/ASET56582.2023.10180891.
52. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
53. He, H.; Bai, Y.; Garcia, E.A.; 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, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969.
54. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In Proceedings of the Advances in Intelligent Computing; Huang, D.S.; Zhang, X.P.; Huang, G.B., Eds., Berlin, Heidelberg, 2005; pp. 878–887. https://doi.org/10.1007/11538059_91.
55. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences* **2018**, *465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056.
56. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigm.* **2011**, *3*, 4–21. https://doi.org/10.1504/IJKESDP.2011.039875.

 doi:10.20944/preprints202510.2132.v1

57. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* **2009**, *36*, 5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108.
58. Hart, P. The condensed nearest neighbor rule (Corresp.). *IEEE Transactions on Information Theory* **1968**, *14*, 515–516. https://doi.org/10.1109/TIT.1968.1054155.
59. Wilson, D.L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics* **1972**, *SMC-2*, 408–421. https://doi.org/10.1109/TSMC.1972.4309137.
60. Tomek, I. An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics* **1976**, *SMC-6*, 448–452. https://doi.org/10.1109/TSMC.1976.4309523.
61. Smith, M.R.; Martinez, T.; Giraud-Carrier, C. An instance level analysis of data complexity. *Machine Learning* **2014**, *95*, 225–256. https://doi.org/10.1007/s10994-013-5422-z.
62. Zhang, J.; Mani, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In Proceedings of the Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets, 2003.
63. Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. In Proceedings of the Artificial Intelligence in Medicine; Quaglini, S.; Barahona, P.; Andreassen, S., Eds., Berlin, Heidelberg, 2001; pp. 63–66. https://doi.org/10.1007/3-540-48229-6_9.
64. Kubat, M. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning, 06 1997, pp. 179–186.
65. Bach, M.; Werner, A.; Palt, M. The Proposal of Undersampling Method for Learning from Imbalanced Datasets. *Procedia Computer Science* **2019**, *159*, 125–134. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019, https://doi.org/10.1016/j.procs.2019.09.167.
66. Tomek, I. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics* **1976**, *SMC-6*, 769–772. https://doi.org/10.1109/TSMC.1976.4309452.
67. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. https://doi.org/10.1145/1007730.1007735.
68. Batista, G.; Bazzan, A.; Monard, M.C. Balancing Training Data for Automated Annotation of Keywords: a Case Study. In Proceedings of the Proc. of Workshop on Bioinformatics, 01 2003, pp. 10–18.
69. Opitz, D.; Maclin, R. An empirical evaluation of bagging and boosting for artificial neural networks. In Proceedings of the Proceedings of International Conference on Neural Networks (ICNN'97), 1997, Vol. 3, pp. 1401–1405 vol.3. https://doi.org/10.1109/ICNN.1997.613999.
70. Chen, C.; Liaw, A.; Breiman, L. Using Random Forest to Learn Imbalanced Data. Technical Report 666, University of California, Berkeley 110 (1-12): 24, 2004.
71. Davis, L., Ed. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold, 1991.
72. Woolson, R.F., Wilcoxon Signed-Rank Test. In *Encyclopedia of Biostatistics*; John Wiley & Sons, Ltd, 2005. https://doi.org/10.1002/0470011815.b2a15177.
73. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **2002**, *6*, 182–197. https://doi.org/10.1109/4235.996017.
74. Blank, J.; Deb, K. Pymoo: Multi-Objective Optimization in Python. *IEEE Access* **2020**, *8*, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567.