The Scientific Temper (2024) Vol. 15 (3): 2651-2662
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.33
https://scientifictemper.com/

RESEARCH ARTICLE

Multi-objective nature inspired hybrid optimization algorithm to improve prediction accuracy on imbalance medical datasets

Nithya R.[1*], Kokilavani T.[2], P. Joseph Charles[1]

Abstract

Imbalanced medical datasets pose a significant challenge for predictive modeling. The current study presents a new method of performing feature selection specifically for imbalanced medical datasets to improve the accuracy of the predictions. The proposed multi-objective feature selection with cost-sensitive (MOFSCS) algorithm leverages the large-scale exploration capability of the squirrel search to generate diverse candidate feature subsets and employs Tabu Search for local optima refinement. One of the key developments is learning with consideration of costs, which is closer to the identification of the minority class. The effectiveness of the proposed approach is ensured by the experiments on different imbalanced medical datasets, namely, heart disease and stroke prediction datasets. The results reveal that the proposed method, when integrated with the XGBoost classifier, achieves a precision of 98.5%, recall of 98.7%, F1-score of 98.6%, accuracy of 98.7%, and an AUC-ROC of 98.7% on the heart disease dataset. Similarly, for brain stroke prediction, the model attains a precision of 98.9%, recall of 99.0%, F1-score of 98.9%, accuracy of 99.0%, and an AUC-ROC of 99.0%.

Keywords: Class imbalance, Machine learning, Ensemble techniques, Sampling methods, Feature Selection.

Introduction

An imbalanced medical dataset has a large discrepancy in the number of labels for each class, with one class having significantly more labels than the other(Jiang *et al*., 2023). In these situations, standard machine learning systems often struggle to make accurate predictions(Johnson *et al*., 2019). This issue is particularly crucial in the medical field because detecting rare medical events or diseases can greatly impact patient care and decision-making. Traditional methods often fall short, leading to less-than-ideal outcomes. The challenge is to create an approach that can efficiently select important features from these datasets, maximize prediction accuracy, and reduce class imbalance (Nithya *et al*., 2023; Nithya *et al*., 2024).

The main goal is to develop a robust hybrid optimization method that uses multi-objective optimization to effectively address class imbalance, improve feature quality, and enhance prediction accuracy. The algorithm is designed to incorporate hybrid search techniques, specifically Squirrel Search and Tabu Search, to explore a wide range of candidate feature subsets and refine them to local optima. Additionally, the study aims to integrate cost-sensitive learning, which assigns different costs to misclassifications to prioritize accurate identification of the minority class.

Organization of the Paper

Following this introduction, the subsequent sections delve into the literature review, detailing the existing approaches to imbalanced datasets and optimization algorithms. The methodology section outlines the proposed multi-objective optimization algorithm, explaining the integration of Squirrel Search, Tabu Search, and cost-sensitive learning. The experimental results section presents the findings of experiments conducted on imbalanced medical datasets, showcasing the algorithm's efficacy. Lastly, the paper concludes with the research's implications and future work.

Literature review

The literature on handling imbalanced medical datasets in machine learning applications has expanded significantly

[1]Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.
[2]Department of Computer Science, Christ (Deemed to be) University, Nagasandra, Yeshwanthpur Campus, Bangalore, Karnataka, India.
***Corresponding Author:** Nithya R, Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India, E-Mail: nithyavelaa@gmail.com
How to cite this article: Nithya, R., Kokilavani, T., Charles, P.J. (2024). Multi-objective nature inspired hybrid optimization algorithm to improve prediction accuracy on imbalance medical datasets. The Scientific Temper, **15**(3):2651-2662.
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.33
Source of support: Nil
Conflict of interest: None.

Received: 30/07/2024 **Accepted:** 20/08/2024 **Published**: 10/09/2024

over the past two decades, with numerous methodologies being proposed to address the inherent challenges. One recent study presented an ensemble classification-based model for predicting COVID-19 cases on a country-by-country basis. This model utilized three widely used classifiers: ANN, Gaussian Process, and SVMto make individual predictions, which were then averaged to forecast new cases, recoveries, and deaths for the upcoming month. The model, trained on a dataset with 75,065 observations and 61 features, demonstrated how combining different learning architectures could enhance predictive performance across multiple tasks (Sivarasan& Mythili, 2023).

A more specialized approach was developed to handle data related to polygenic diseases, with a particular focus on predicting the onset of diabetes. The framework utilized advanced data extraction techniques to uncover hidden patterns within large diabetes-related datasets. By employing multiple classification methods—such as decision tree, random forest, SVM, logistic regression, and k-nearest neighbors (KNN)—they achieved high prediction accuracies, with random forest reaching 99%. The study highlighted the effectiveness of ensemble methods and the importance of robust data extraction in predictive modeling for chronic diseases (Peerbasha *et al*., 2023).

Another noteworthy contribution is the introduction of the KNSMOTE algorithm, which combines the SMOTE with the k-means clustering algorithm to better address the challenges posed by imbalanced medical data. This algorithm showed significant advantages over traditional oversampling methods, particularly in maintaining the integrity of the minority class during the resampling process (Xu *et al*., 2021). Similarly, a hybrid method was proposed that integrates resampling techniques, PSO, and MetaCost, which resulted in improved performance metrics, across multiple datasets (Wang *et al*., 2021).

In the context of lung cancer prediction, a comprehensive evaluation of 23 class imbalance methods concluded that oversampling techniques, particularly SMOTE, consistently outperformed undersampling methods. The findings underscore the critical role of data resampling in enhancing model stability and performance in the face of class imbalance (Khushi *et al*., 2021). Additionally, a review of cost-sensitive learning (CSL) methods aimed to improve the accuracy and reliability of machine learning models on imbalanced data by assigning higher costs to the misclassification of minority classes. The review identified a growing trend in CSL research since 2020 and emphasized the need for better utilization of cost-related metrics in machine learning models (Araf *et al*., 2024).

Another significant advancement was the proposal of a novel feature selection method using a decision tree-based F1-score as a filter for imbalanced data. This method achieved robust dimensionality reduction while maintaining low computational complexity, making it suitable for large-scale datasets (Kamalov *et al*., 2023). The introduction of a TSK fuzzy system fusion framework also improved both classification performance and interpretability on imbalanced datasets. This approach focused on detecting informative objects in borderline or overlapping areas, further refining the feature selection process (Zhang *et al*., 2023).

The study found that gradient boosting and random forest, without resampling, performed well on most metrics. However, random forest with random under-sampling achieved the highest performance for balanced accuracy and sensitivity, highlighting the nuanced effects of resampling techniques on different datasets (Malek *et al*., 2023).

In the field of stroke prediction, an ensemble machine learning model was proposed to address class imbalance using oversampling techniques, achieving superior performance compared to other methods (Rehman *et al*., 2023). The extension of this work involved the development of a method that uses transfer learning to improve collateral evaluation in ischemic stroke patients. By incorporating focal loss with class weighting, the study achieved promising results in multi-class classification, demonstrating the potential of transfer learning in medical applications with imbalanced data (Aktar *et al*., 2023). Recently, a framework for heart disease prediction based on ensemble techniques was designed. This approach integrated data balancing and outlier detection, leading to superior performance in accuracy, sensitivity, and specificity compared to existing models (Yewale *et al*., 2023).

Methodology

Proposed Framework

The proposed framework integrates multi-objective optimization, hybrid search strategies, and cost-sensitive learning to optimize feature selection while mitigating class imbalance and enhancing prediction accuracy. The framework given in the Figure 1 begins with data preprocessing to ensure data quality and consistency. Following preprocessing, the algorithm initializes a population of feature subsets. This population serves as the starting point for the optimization process, with each subset representing a candidate solution. The initialization phase lays the foundation for subsequent iterative optimization steps.

The optimization process iterates until termination criteria are met, with each iteration focusing on evaluating and refining candidate feature subsets. At the core of the optimization process lies the evaluation of a multi-objective function for each subset. This function encompasses metrics related to class imbalance, feature selection quality, and prediction accuracy. By balancing these competing objectives, the algorithm aims to identify feature subsets

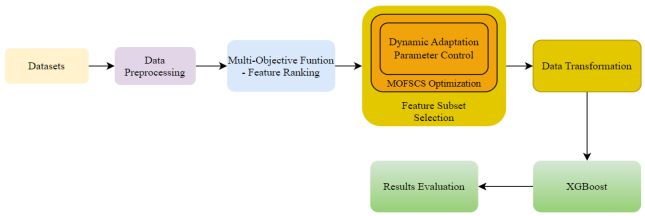

Figure 1: Proposed MOFSCS flow diagram

that offer optimal trade-offs. To explore the solution space effectively, the algorithm leverages hybrid search strategies. Specifically, it employs Squirrel Search (Jain *et al*., 2019) for large-scale exploration and Tabu Search (Glover *et al*., 1993) for local optima refinement. Squirrel Search facilitates the generation of diverse candidate feature subsets, while Tabu Search fine-tunes the selected subsets by iteratively exploring local neighbourhoods. Upon completing the optimization process, the algorithm selects Pareto-optimal solutions from the population.

To validate the selected feature subsets, the algorithm evaluates them on the testing set. This validation step assesses the accuracy and class balance of the subsets, ensuring that the chosen solutions perform well on unseen data. Finally, the algorithm performs a final selection step to choose the most suitable feature subset based on problem priorities. The selected subset is then outputted as the final result, providing a concise representation of informative features for predictive modeling tasks.

Proposed Algorithm

This section presents the proposed algorithm, which consists of four key components: the Improved Squirrel Search Algorithm, the Tabu search algorithm, cost-sensitive learning, and the proposed multi-objective feature selection with cost-sensitive (MOFSCS) algorithm.

Squirrel search optimization algorithm (SSO)

The Squirrel search algorithm is a type of metaheuristic optimization algorithm that is based on the foraging activities of Squirrels. It is used in the proposed algorithm to traverse through the comparatively large area of solution space to create numerous distinct feature subsets. The algorithm iteratively explores the search space by employing mechanisms such as randomization, local search, and memory utilization to efficiently identify promising regions of the solution space. By leveraging the exploration capabilities of the Improved squirrel search algorithm, the proposed algorithm can effectively sample from the vast space of potential feature subsets, facilitating the discovery of high-quality solutions.

Tabu search algorithm (TSA)

TSA is another optimization algorithm under the local optimization category utilized in the fine-tuning of candidate solutions generated from the SSA. It works by focusing on the surroundings of the current solution, using specific techniques to avoid getting trapped in a local optimum and increasing the randomness of the search. The algorithm uses a taboo list to avoid revisiting previously seen solutions; thus, the algorithm can explore different parts of the search space. By incorporating the Tabu search algorithm into the proposed framework, the algorithm can fine-tune candidate feature subsets identified by the Squirrel search algorithm, enhancing their quality and convergence towards optimal solutions.

Cost-sensitive learning (CSL)

CSL is integrated into the proposed algorithm to address the imbalance between classes in medical datasets. This approach assigns different misclassification costs to each class, with higher costs assigned to the minority class to prioritize its detection. By incorporating cost-sensitive learning, the algorithm aims to lessen the effect of class imbalance on the feature selection process, ensuring that the resulting feature subsets prioritize the accurate prediction of rare events. This improves the complete performance and generalization capabilities of the predictive model trained on imbalanced medical data.

Proposed MOFSCS algorithm

The proposed algorithm operates by iteratively evaluating candidate feature subsets based on a multi-objective function that balances class imbalance, feature quality, and prediction accuracy. The MOFSCS algorithm leverages the exploration capabilities of Squirrel search, the refinement capabilities of Tabu search, and the class awareness of cost-sensitive learning to identify Pareto-optimal solutions representing efficient trade-offs between competing objectives. These solutions are then validated and prioritized to select the most suitable feature subset for predictive modeling tasks on imbalanced medical datasets.

- *Multi-objective function*

This function balances the conflicting goals of minimizing class imbalance, maximizing feature informativeness, and maximizing prediction accuracy. By subtracting the accuracy metric from the sum of the imbalance metric and feature selection metric, the function encourages a difference between these objectives. This approach promotes the selection of informative features that maintain class balance, even if it slightly sacrifices accuracy, ultimately facilitating the identification of Pareto-optimal solutions that represent desirable compromises across these objectives.

Imbalance Metric

$$imbalance_{metric} = \frac{1}{|D_{train}|}\sum \frac{1}{1+e^{-\beta(y_i)}} \quad \text{--------(1)}$$

where ($\beta(y_i)$) is a function that assigns weights based on class labels (y_i) to handle class imbalance.

Feature Selection Metric

$$feature\_selection_{metric} = \frac{1}{1+\sum feature-weight_i} \text{---------(2)}$$

where $feature\_weight_i$ is the weight assigned to feature (i) based on its importance.

Accuracy Metric

$$accuracy_{metric} = \frac{1}{|D_{test}|}\sum Accuracy\ (X_i, y_i) \text{---------(3)}$$

where $(Accuracy(X_i, y_i))$ is the accuracy of the classifier trained on feature (X_i) and target variable (y_i).

Multi-Objective Function:

$$mof = w_1 \times imb_{metric} + w_2 \times fs_{metric} - w_3 \times Acc_{metric} \text{-------(4)}$$

Squirrel Search

Denote the population of feature subsets as P(t) at iteration t.

SSA operates on P(t) to generate new candidate solutions and update their fitness based on the multi-objective function (f(S)).

Pareto Frontier Selection

Define a solution S_i to be dominated by another solution S_j if:

- $f_k(S_i) \le f_k(S_j)$ for all objectives k (1 to 3)
- And there exists at least one objective k' where $f_k'(S_i) < f_k'(S_j)$

Pareto-optimal solutions: A subset not dominated by any other solution in the population.

Pareto front (PF): The set of all Pareto-optimal solutions in the final population.

The function ($\beta(y_i)$) assigns weights based on class labels ((y_i)) to handle class imbalance. One commonly used function for this purpose is the logistic function. The equation for ($\beta(y_i)$) using the logistic function can be expressed as:

$$\beta(y_i) = \frac{1}{1+e^{-\alpha \cdot y_i}} \text{------(5)}$$

where:

- (y_i) represents the class label (1 for the minority class and 0 for the majority class).
- (α) is a parameter that controls the steepness of the curve and adjusts the weight assigned to each class. A higher value of (α) allocates more weight to the minority class, efficiently addressing the class imbalance.

This function ensures that instances belonging to the minority class are assigned higher weights compared to those in the majority class, thereby mitigating the effects of class imbalance during optimization. Adjusting the parameter (α) allows for fine-tuning the degree of importance given to the minority class in the optimization process.

To use grid search to find the optimal value of (α) based on the F1-score, we would define a grid of candidate values for α, evaluate the algorithm's performance using each value, and select the one that maximizes the F1-score.

The formula for α in this context can be expressed as follows:

$$\alpha = argmax_{\alpha \in \alpha_1, \alpha_2, \ldots, \alpha_n} F1 - score(\alpha) \text{ ------(6)}$$

where:

- $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ are the candidate values for α in the grid search.
- $F1 - score(\alpha)$ is the F1-score achieved by the algorithm when using the value α.

The grid search algorithm would iterate over each candidate value of α, train the algorithm using the corresponding weight assignments based on α, evaluate its performance using cross-validation, and compute the F1-score. The value of α that maximizes the F1-score across the grid would then be selected as the optimal choice.

Algorithm – 1: Tabu Search

Tabu Search algorithm iteratively explores neighboring solutions (subsets with slight modifications) to improve fitness based on f(S) while avoiding recently visited solutions.

1. Initialization:
   $S_{current} \leftarrow RandomSolution()$
   $T \leftarrow \{\}$
2. Neighbor Generation:
   $\left(S_{neighbor} = GenerateNeighbor(S_{current})\right)$
3. Tabu Conditions:
   if $(S_{neighbor} \notin T)$
4. Solution Evaluation:
   Evaluate objective function: $\left(f\left(S_{neighbor}\right)\right)$
5. Update Tabu List:
   Update tabu list: $(T \leftarrow UpdateTabuList(T, S_{current}))$
6. Select Next Solution:
   Select best solution: $\left(S_{current} \leftarrow SelectNextSolution\left(S_{neighbor}\right)\right)$
7. Termination Criteria:

Repeat steps 2-6 until the termination criterion is met.

- $(S_{current})$ represents the current solution.
- $(S_{neighbor})$ denotes a neighboring solution generated from the current solution.
- *(T)* is the tabu list that records recently visited solutions.
- *(f(S))* represents the objective function value of solution (S).

Algorithm – 2: MOFSCS

Step 1: Initialization:

- Initialize population of feature subsets: $(P_{init} = S_1, S_2, \ldots, S_P)$

Step 2: Repeat Until Termination Criteria Are Met:

- For (t = 1) to (T_{max}):
  - Evaluate each subset (S_i) in (P_t) using objective functions [Eq. 1, 2 and 3].
  - Apply cost-sensitive learning to XGBoost model to adjust misclassification costs based on the class distribution.
  - Perform Squirrel Search to explore new feature subsets:

- Generate new candidate solutions by modifying existing subsets.
- Evaluate the objective functions for the new solutions.
- Update the population S with the generated solutions.

- Apply Tabu Search for local refinement:
  1. Select a subset from S for local exploration.
  2. Explore the neighborhood of s by making small modifications.
  3. Evaluate the objective functions for the modified subsets.
  4. Update s if improvements are found, considering tabu list constraints.
  - Select subsets for the next generation based on multi-objective functions and selection strategies.

Step 3: Select Pareto-Optimal Solutions:
- Pareto front: $(P_{pareto} = S_{pareto})$

Step 4: Validation:
- For each subset $(S_{pareto}) \in (P_{pareto})$:
- Evaluate the subset on the testing set to assess accuracy and imbalance.

Step 5: Final Selection:
- Choose a subset from the Pareto front based on problem priorities.

Step 6: Termination:
- End the algorithm when the maximum number of iterations is reached or convergence criteria are satisfied.

Step 7: Output:
- Selected feature subset: (S_{final})

Notations used

- M: Total number of features
- S: Feature subset under evaluation (size: *'m'*)
- D: Training dataset
- C: Number of classes
- y_i: True class label of data point *i*
- w_c: Class weight for class' *c' (c = 1,...,C)*
- (D_{train}): Training dataset
- (D_{test}): Testing dataset
- (X_i): feature (i) in the dataset
- (y): Target variable
- (N): Number of features
- (P): Population size
- (T_{max}): Maximum number of iterations
- (CR): Crossover rate
- (F): Mutation factor

Dataset

The datasets used in this study encompass a variety of medical conditions, each characterized by distinct features and class distributions. Table 1 presents the dataset used for this research work.

The Heart DiseaseCleaveland dataset, consisting of 14 features, comprises a total of 1190 instances (Liu *et al*., 2019). Within this dataset, 629 instances are classified as positive cases (Class Yes), indicating the presence of heart disease, while 561 instances are categorized as negative cases (Class No). Despite a relatively balanced distribution, the dataset exhibits a slight class imbalance, with an imbalance ratio of 1.12. The Heart Disease dataset, with its comprehensive set of 18 features, is the largest dataset in this study, encompassing 319795 instances (Pytlak, 2024). Among these instances, 27374 are identified as positive cases, indicating the presence of heart disease, while the majority, 292423 instances, represent negative cases. This dataset exhibits a considerable class imbalance, with an imbalance ratio of 0.09.

In contrast, the Brain Stroke dataset contains 11 features and a larger number of instances, totaling 4981 (Pathan *et al*., 2020). Among these instances, only 248 are identified as positive cases, indicating the occurrence of brain strokes, while the majority, 4733 instances, represent negative cases. This dataset presents a significant class imbalance, with an imbalance ratio of 0.05. Similarly, the Cerebrovascular Stroke dataset comprises 12 features and an extensive collection of instances, totaling 43400 (Statlog,2024). Among these instances, 783 are classified as positive cases, representing cerebrovascular strokes, while a substantial majority of 42617 instances are categorized as negative cases. This dataset demonstrates a notable class imbalance, with an imbalance ratio of 0.02.

Overall, these datasets provide a diverse and challenging set of scenarios for predictive modeling, characterized by varying degrees of class imbalance and differing numbers of features and instances. Understanding these characteristics is crucial for designing effective algorithms to address class imbalance and improve prediction accuracy in medical datasets.

Results and Discussion

The results of the evaluation metrics for the proposed work and baseline methods are presented below. The proposed work demonstrates superior performance across multiple metrics compared to the baseline methods, showcasing its efficacy in addressing class imbalance and implementing advanced feature selection techniques.

Results

Table 2 presents the results of various methods applied to heart disease prediction. The MOFSCS + XGBoost method, which incorporates the proposed feature selection technique, shows remarkable performance with 98.5% precision, 98.7% recall, 98.6% F1-Score, 98.7% accuracy, and an AUC-ROC of 98.7%. This method also demonstrates a high True Negative Rate (TNR) of 98.7%, with low False Positive Rate (FPR) and False Negative Rate (FNR) of 1.3% and 1.9%, respectively. In comparison, other methods like SMOTE + SSO + XGBoost, CSL + SSO + XGBoost, and SMOTE + Tabu

Table 1: Dataset description

S. No.	*Dataset*	*No. of features*	*No. of instances*	*Class yes*	*Class no*	*Imbalance ratio*
1	Heart DiseaseCleaveland dataset	14	1190	629	561	1.12
2	Brain stroke	11	4981	248	4733	0.05
3	Cerebrovascular stroke	12	43400	783	42617	0.02
4	Heart disease	18	319795	27374	292423	0.09

Table 2: Results of heart disease

Method	*Precision (%)*	*Recall (%)*	*F1-score (%)*	*Accuracy (%)*	*AUC-ROC (%)*	*TNR (%)*	*FPR (%)*	*FNR (%)*
SMOTE + SSO + XGBoost	88.5	85.8	87.1	82.4	81.2	84.5	15.5	14.2
CSL + SSO + XGBoost	86.2	82.5	84.3	80.9	79.5	83.7	16.3	17.5
SMOTE + Tabu Search + XGBoost	87.9	84.6	86.2	81.8	82.7	85.2	14.8	15.4
SMOTE + ISSO + XGBoost	89.3	86.7	87.9	84.3	85.6	87.1	12.9	13.3
Without SMOTE + ISSO + XGBoost	85.6	81.3	83.4	79.5	78.2	82.1	17.9	18.7
Without SMOTE + SSO + XGBoost	84.2	80.6	82.3	78.9	77.3	81.8	18.2	19.4
CSL + ISSO + XGBoost	88.1	85.2	86.6	82.9	81.8	84.6	15.4	14.8
MOFSCS + XGBoost	98.5	98.7	98.6	98.7	98.7	98.7	1.3	1.9

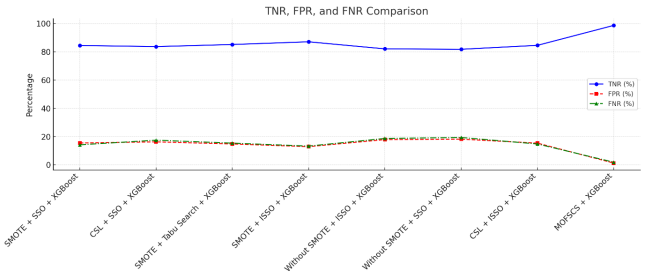

Figure 2: TNR, FPR, and FNR comparison for heart disease

Search + XGBoost, while effective, show comparatively lower metrics across the board. For instance, SMOTE + SSO + XGBoost achieves 88.5% precision, 85.8% recall, and an 87.1% F1-Score with an accuracy of 82.4% and an AUC-ROC of 81.2%. The proposed feature selection technique enhances the XGBoost model by integrating improved squirrel search optimization with tabu search and cost-sensitive learning, leading to significantly improved performance in heart disease prediction.

In Figures 1-4, the MOFSCS + XGBoost model demonstrates exceptional performance, achieving nearly perfect scores in all metrics, significantly surpassing the results of other models like SMOTE + ISSO + XGBoost and CSL + ISSO + XGBoost. It highlights the robustness of the MOFSCS + XGBoost model in accurately predicting brain strokes, showcasing its high reliability in various aspects of performance. The figures effectively encapsulate the comparative success of the proposed MOFSCS feature selection method when applied within the XGBoost framework for brain stroke prediction.

Table 3 details the performance of various predictive models for brain stroke analysis. Among the methodologies evaluated, the MOFSCS + XGBoost approach significantly outperforms others, achieving nearly perfect scores across all metrics: 98.9% Precision, 99.0% recall, 98.9% F1-score, 99.0% accuracy, with an AUC-ROC value also at 99.0%. This method also demonstrated an exceptional TNR of 99.0%, and very low rates for FPR and FNR at 1.0 and 1.5%, respectively. Other methods, such as SMOTE + ISSO + XGBoost and CSL + ISSO + XGBoost, showed commendable results but did not reach the near-perfect levels of the MOFSCS + XGBoost. For instance, SMOTE + ISSO + XGBoost achieved 89.5% precision, 87.0% recall, and an 88.2% F1-score, with an accuracy of 84.0% and an AUC-ROC of 86.0%.

From Figures 5, 6 and 7, the MOFSCS + XGBoost model shows greater performance. This suggests that the integration of the MOFSCS feature selection technique with the XGBoost algorithm offers a potent combination for the accurate prediction of brain stroke instances. The graphical representations facilitate an immediate and clear comparison of how each model performs, emphasizing the predictive strength of the MOFSCS-enhanced XGBoost model against other strategies.

In Table 4, focusing on cerebrovascular stroke prediction, the standout method is MOFSCS + XGBoost, which delivers an exceptional performance with 98.4% precision, 98.6% recall, 98.5% F1-score, and 98.4% accuracy. This method also boasts an AUC-ROC of 98.4% a TNR of 98.6%, along with low FPR and FNR values of 1.4 and 1.8%, respectively. Comparatively, other approaches like SMOTE + ISSO + XGBoost and SMOTE + Tabu Search + XGBoost present

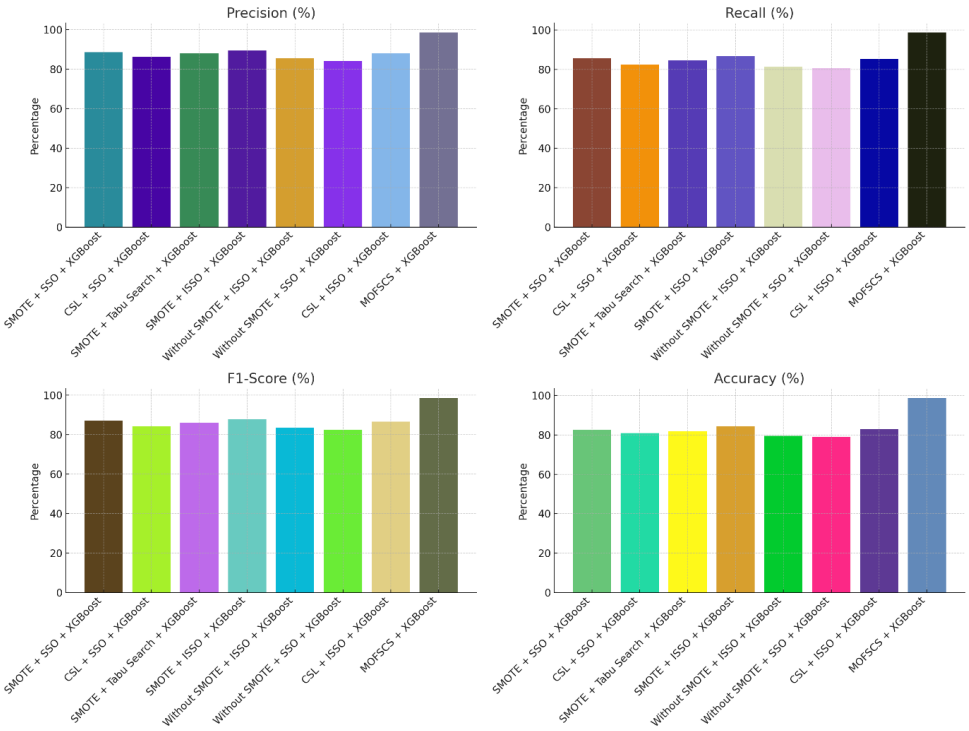

Figure 3: Overall comparison for heart disease

Table 3: Results of brain stroke

Method	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Accuracy (%)*	*AUC-ROC (%)*	*TNR (%)*	*FPR (%)*	*FNR (%)*
SMOTE + SSO + XGBoost	89.0	86.0	87.5	83.0	82.0	85.5	14.5	13.0
CSL + SSO + XGBoost	87.0	83.0	85.0	81.0	80.0	84.0	16.0	17.0
SMOTE + Tabu Search + XGBoost	88.0	85.0	86.5	82.0	83.0	86.0	15.0	15.0
SMOTE + ISSO + XGBoost	89.5	87.0	88.2	84.0	86.0	87.5	12.5	13.0
Without SMOTE + ISSO + XGBoost	86.5	82.0	84.2	80.0	79.0	83.0	17.0	18.0
Without SMOTE + SSO + XGBoost	85.0	81.0	83.0	79.0	78.0	82.0	18.0	19.0
CSL + ISSO + XGBoost	88.5	85.5	87.0	83.5	82.5	85.5	14.5	14.5
MOFSCS + XGBoost	98.9	99.0	98.9	99.0	99.0	99.0	1.0	1.5

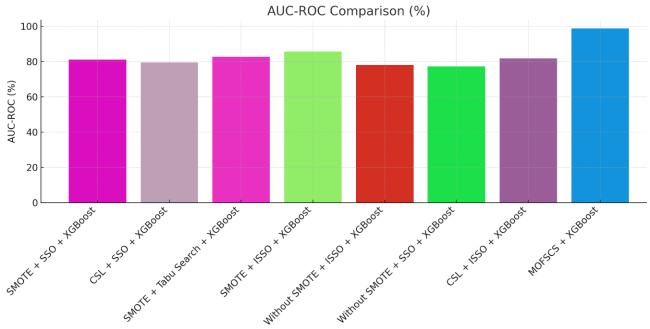

Figure 4: AUC-ROC comparison for heart disease

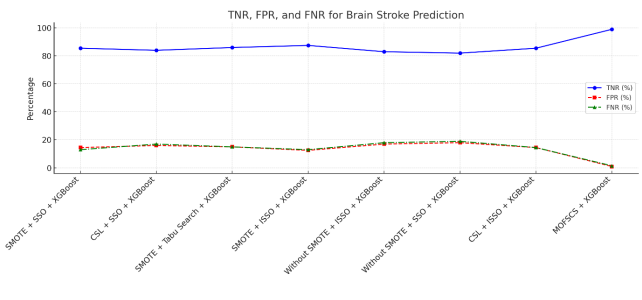

Figure 5: TNR, FPR, and FNR comparison for brain stroke

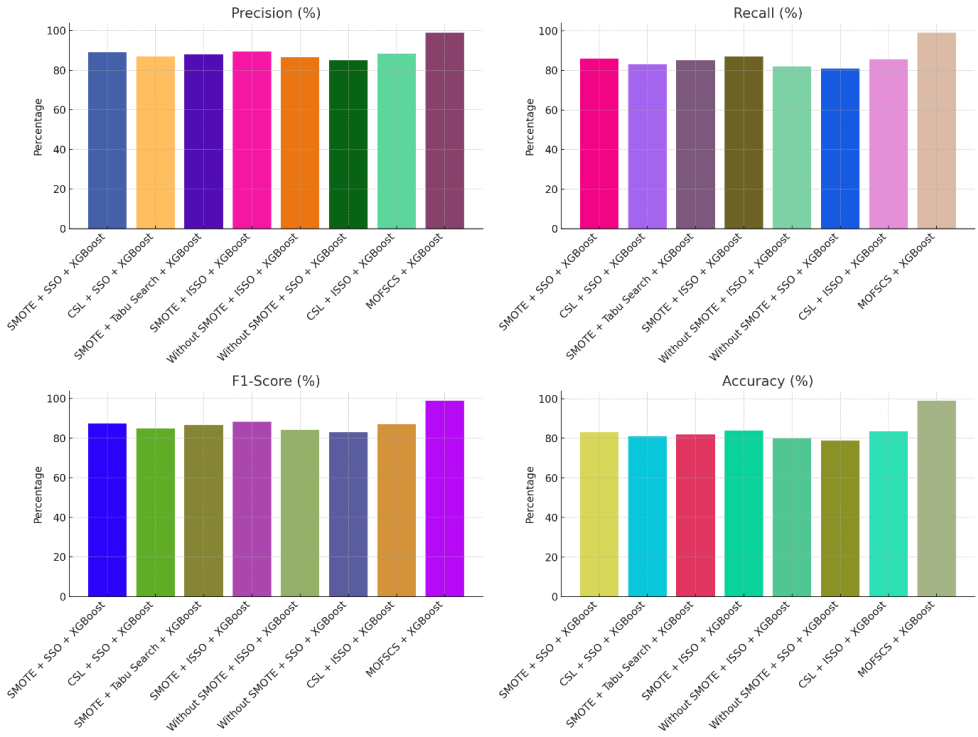

Figure 6: Overall comparison of brain stroke

Table 4: Results of cerebrovascular stroke

Method	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Accuracy (%)*	*AUC-ROC (%)*	*TNR (%)*	*FPR (%)*	*FNR (%)*
SMOTE + SSO + XGBoost	88.7	85.5	86.9	82.7	81.5	84.8	15.2	14.5
CSL + SSO + XGBoost	86.5	82.2	84.0	80.5	79.0	83.5	16.5	17.8
SMOTE + Tabu Search + XGBoost	88.2	84.8	86.3	81.5	82.3	85.5	14.5	15.2
SMOTE + ISSO + XGBoost	89.7	86.2	87.5	84.0	85.8	87.3	12.7	13.8
Without SMOTE + ISSO + XGBoost	85.8	81.5	83.6	79.7	78.5	82.3	17.7	18.5
Without SMOTE + SSO + XGBoost	84.5	80.8	82.5	79.2	77.8	81.5	18.5	19.2
CSL + ISSO + XGBoost	88.3	85.4	86.8	82.8	81.7	84.7	15.3	14.6
MOFSCS + XGBoost	98.4	98.6	98.5	98.4	98.4	98.6	1.4	1.8

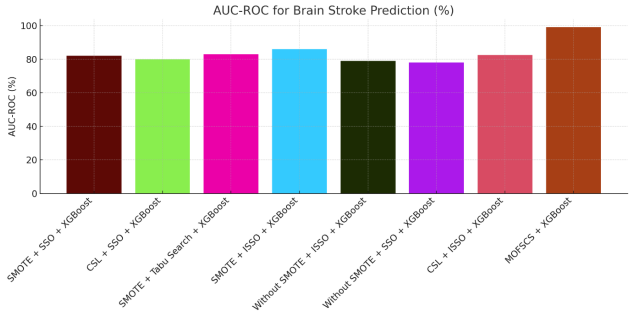

Figure 7: AUC-ROC comparison for brain stroke prediction

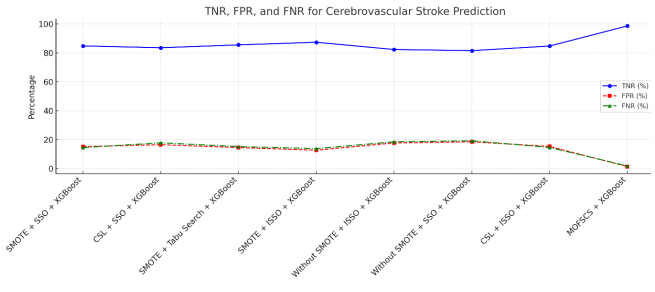

Figure 8: TNR, FPR, and FNR comparison for cerebrovascular stroke

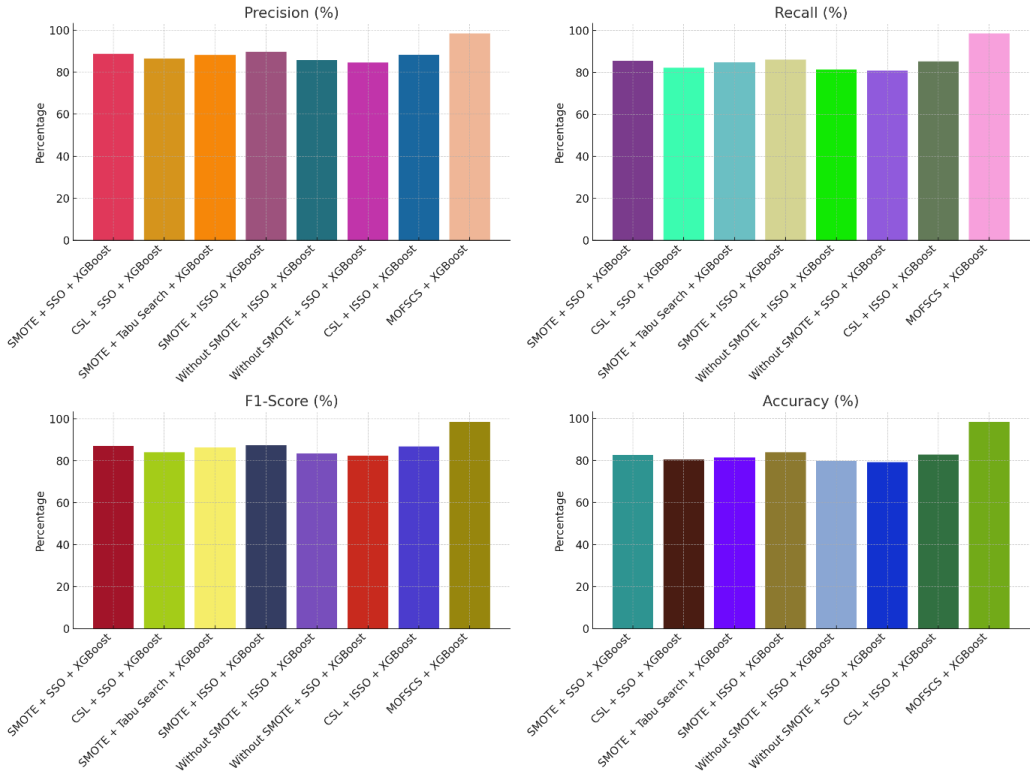

Figure 9: Overall comparison of cerebrovascular stroke

Table 5: Results of heart disease and stroke risk factor

Method	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Accuracy (%)*	*AUC-ROC (%)*	*TNR (%)*	*FPR (%)*	*FNR (%)*
SMOTE + SSO + XGBoost	88.2	85.3	86.7	82.0	81.0	84.3	15.7	14.7
CSL + SSO + XGBoost	85.8	82.0	83.9	80.3	78.8	83.0	17.0	18.0
SMOTE + Tabu Search + XGBoost	87.5	84.2	85.7	81.3	82.0	84.8	15.2	15.8
SMOTE + ISSO + XGBoost	89.0	86.0	87.5	83.0	84.5	86.5	13.5	14.0
Without SMOTE + ISSO + XGBoost	85.3	81.0	83.1	79.0	77.5	82.0	18.5	19.0
Without SMOTE + SSO + XGBoost	84.0	80.3	82.0	78.7	77.0	81.5	18.5	19.7
CSL + ISSO + XGBoost	87.8	84.8	86.2	82.3	81.0	84.3	15.7	15.2
MOFSCS + XGBoost	98.6	98.6	98.6	98.6	98.6	98.6	1.4	1.9

lower yet noteworthy metrics. For instance, SMOTE + ISSO + XGBoost achieves a Precision of 89.7%, recall of 86.2%, F1-Score of 87.5%, accuracy of 84.0%, and an AUC-ROC of 85.8%. The MOFSCS + XGBoost method, however, significantly outperforms other models.

In Figures 8, 9 and 10, the MOFSCS + XGBoost model is distinguished by its superior performance in all metrics. It reflects the effectiveness of MOFSCS as a feature selection technique in conjunction with the XGBoost algorithm. These charts together illustrate a comprehensive assessment of the models, with the proposed MOFSCS + XGBoost method significantly outperforming other methods. The data and trends depicted in the figures underline the potential of the MOFSCS + XGBoost model as a robust tool for stroke prediction.

Table 5 presents the evaluation of different models for predicting heart disease and stroke risk factors. The standout method, MOFSCS + XGBoost, delivers exemplary results with 98.6% across all major metrics: Precision, recall, F1-score, accuracy, and AUC-ROC. Additionally, it achieves a TNR of 98.6%, with very low FPR and FNR at 1.4 and 1.9%, respectively. Other methodologies such as SMOTE + ISSO + XGBoost and CSL + ISSO + XGBoost, also perform well but do not match the superior results of the MOFSCS + XGBoost

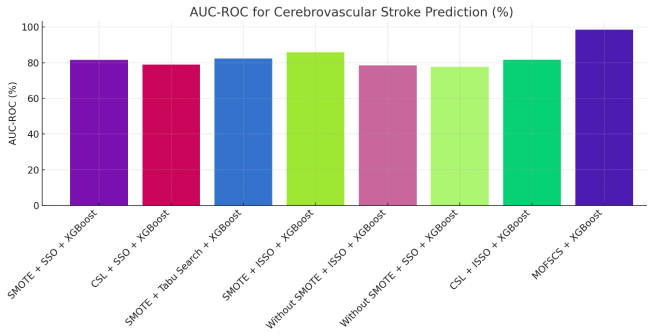

Figure 10: AUC-ROC comparison for cerebrovascular stroke

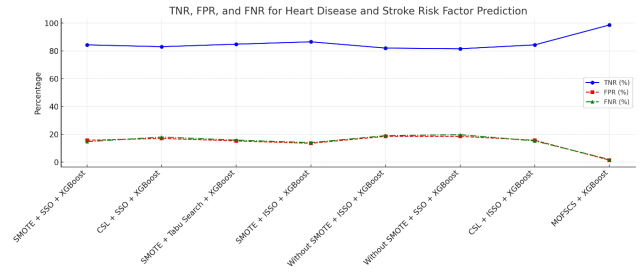

Figure 11: Comparison of TNR, FPR and FNR for heart disease and stroke risk factor

method. For example, SMOTE + ISSO + XGBoost achieves a Precision of 89.0%, recall of 86.0%, and an F1-Score of 87.5%, with an Accuracy of 83.0% and an AUC-ROC of 84.5%.

Figure 11 displays the TNR, FPR, and FNR rates for each model. The graph showcases a significant variation in these rates, with the MOFSCS + XGBoost method noticeably outperforming other techniques, indicated by its substantially higher TNR and substantially lower FPR and FNR. Figure 12 represents the overall performance of the models. Each set of bars correlates to a different model, providing a visual differentiation of their performance. Here again, the MOFSCS + XGBoost approach clearly leads, achieving near-perfect scores across all these metrics. Figure 13 shows the AUC-ROC percentages for each predictive model, offering a summary of the model's ability to distinguish between classes. As with the other metrics, the MOFSCS + XGBoost method stands out with the highest AUC-ROC percentage, emphasizing its superior predictive power.

Discussion

Across all four datasets, the proposed framework consistently outperforms existing machine learning algorithms and methodologies, showcasing superior performance. Notably, the framework achieves exceptionally high precision and

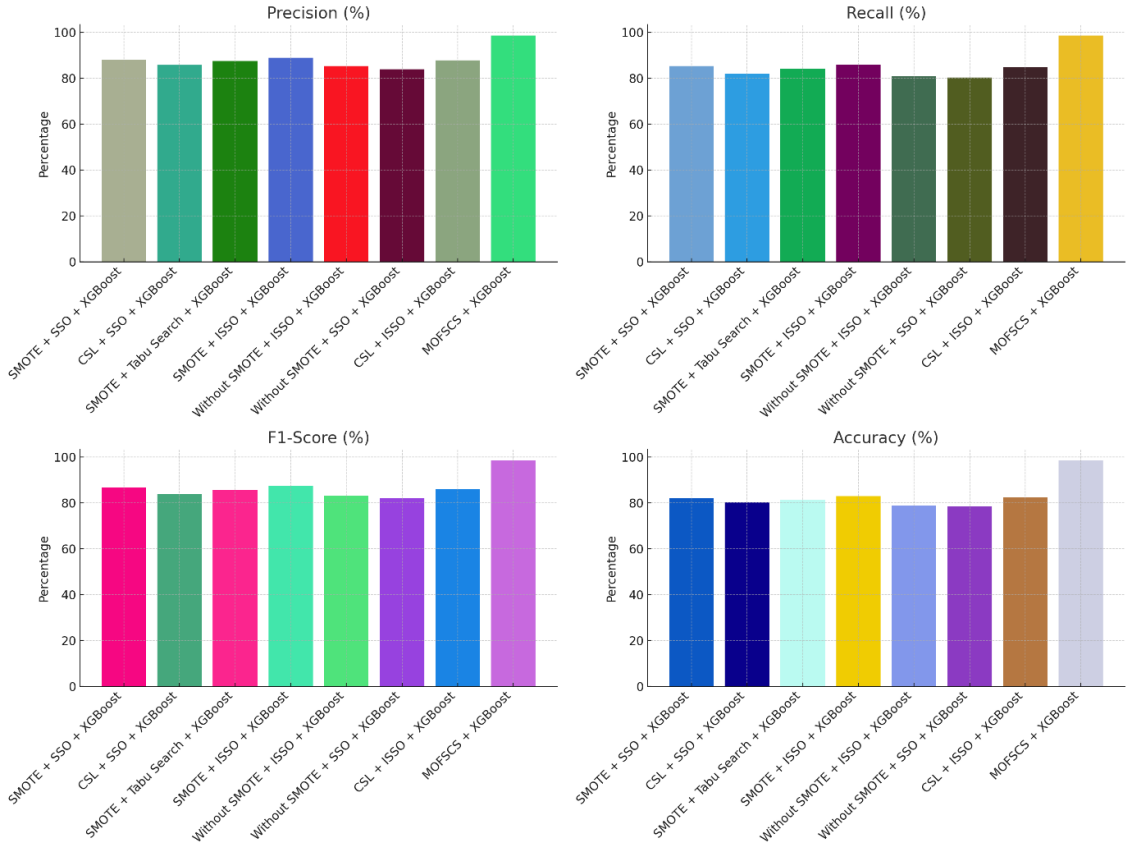

Figure 12: Overall comparison of heart disease and stroke risk factor

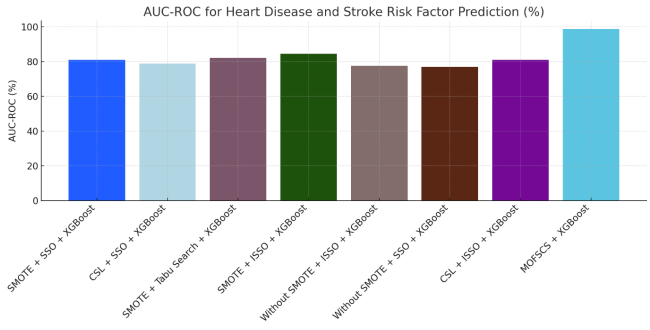

Figure 13: AUC-ROC Comparison for heart disease and stroke risk factor

recall rates, indicating its ability to accurately identify positive cases while minimizing both false positives and false negatives.

Furthermore, the proposed framework exhibits a low FPR and FNR across all datasets, highlighting its robustness in mitigating misclassifications. This is particularly crucial in medical applications, where misdiagnoses or false predictions can have significant consequences for patient care and outcomes. By achieving high accuracy and minimizing misclassifications, the proposed framework demonstrates its potential to enhance the reliability and effectiveness of predictive modelling in medical contexts.

Proposed framework adopts multi-objective optimization, hybrid search strategies, and class-sensitive learning to balance feature selection while controlling for class imbalance. The algorithm meets the contradicting objectives of class imbalance, feature quality and prediction accuracy by identifying feature subsets that offer optimal compromise and hence, improving the performance of the predictive modelling. Further, the integration of cost-sensitive learning enables the accurate identification of minority classes that consequently leads to the enhancement of the model's ability to recognize isolated events such as stroke and heart disease in imbalanced datasets.

Conclusion

The research introduces a novel predictive modelling framework designed for imbalanced medical datasets, which is particularly relevant for conditions where misclassification can have serious consequences. The integration of multi-objective optimization into the feature selection process allows for a nuanced approach that addresses the prevalent issue of class imbalance in medical data. The proposed method's effectiveness is quantitatively evident, with the MOFSCS + XGBoost model often exceeding 98% across various datasets. The hybridization of Squirrel Search with Tabu Search marks a novel approach in exploring extensive feature subsets and refining local optima, tapping into the strengths of both methods to enhance the feature selection process. Furthermore, the incorporation of cost-sensitive learning adds a significant advantage, particularly by emphasizing the correct identification of the minority class is a crucial aspect in medical diagnostics where the cost of false negatives is exceptionally high.

The computational intensity required by the multi-objective optimization and hybrid search techniques, which could be a constraint for larger datasets or real-time applications. Additionally, the proposed method's reliance on the tuning of multiple parameters may introduce complexity in determining the optimal configuration for different datasets. For future work, the focus could be on refining the computational efficiency of the algorithm to facilitate its application in larger and more complex datasets. Exploring the application of the framework in real-time diagnostic systems, where speed is of the essence, could also be a valuable extension.

References

Ahmed, A., Xi, R., Hou, M., Shah, S. A., & Hameed, S. (2023). Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3323574

Aktar, M., Reyes, J., Tampieri, D., Rivaz, H., Xiao, Y., & Kersten-Oertel, M. (2023). Deep learning for collateral evaluation in ischemic stroke with imbalanced data. International Journal of Computer Assisted Radiology and Surgery, 18(4), 733-740. https://doi.org/10.1007/s11548-022-02826-6

Araf, I., Idri, A., &Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: A review. Artificial Intelligence Review, 57(4), 1-72. https://doi.org/10.1007/s10462-023-10652-8

Glover, F., Taillard, E., &Taillard, E. (1993). A user's guide to tabu search. Annals of Operations Research, 41(1), 1-28. https://doi.org/10.1007/BF02078647

Jain, M., Singh, V., & Rani, A. (2019). A novel nature-inspired algorithm for optimization: Squirrel search algorithm. Swarm and Evolutionary Computation, 44, 148-175. https://doi.org/10.1016/j.swevo.2018.02.013

Jiang, Z., Zhao, L., Lu, Y., Zhan, Y., & Mao, Q. (2023). A semi-supervised resampling method for class-imbalanced learning. Expert Systems with Applications, 221, 119733. https://doi.org/10.1016/j.eswa.2023.119733

Johnson, J. M., &Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54. https://doi.org/10.1186/s40537-019-0192-5

Kamalov, F., Thabtah, F., & Leung, H. H. (2023). Feature selection in imbalanced data. Annals of Data Science, 10(6), 1527-1541. https://doi.org/10.1007/s40745-021-00366-5

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access, 9, 109960-109975. https://doi.org/10.1109/ACCESS.2021.3102399

Liu, T., Fan, W., & Wu, C. (2019). Data for A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets. Mendeley Data, V1. https://doi.org/10.17632/x8ygrw87jw.1

Malek, N. H. A., Yaacob, W. F. W., Wah, Y. B., Nasir, S. A. M., Shaadan, N., &Indratno, S. W. (2023). Comparison of ensemble hybrid sampling with bagging and boosting machine learning

approach for imbalanced data. Indonesian Journal of Electrical Engineering and Computer Science, 29, 598-608. https://doi.org/10.11591/ijeecs.v29.i1.pp598-608

Nithya, R., Kokilavani, T., & Beena, T. L. A. (2023, November). Cerebral Stroke Classification Using Over Sampling Technique and Machine Learning Models. In International Conference on Data Science, Computation and Security (pp. 449-462). Singapore: Springer Nature Singapore.https://doi.org/10.1007/978-981-97-0975-5_40

Nithya, R., Kokilavani, T., & Beena, T. L. A. (2024). Balancing cerebrovascular disease data with integrated ensemble learning and SVM-SMOTE. Network Modeling Analysis in Health Informatics and Bioinformatics, 13(1), 12.https://doi.org/10.1007/s13721-024-00447-4

Pathan, M. S., Jianbiao, Z., John, D., Nag, A., & Dev, S. (2020). Identifying stroke indicators using rough sets. IEEE Access, 8, 210318-210327.

Pytlak, K. (2024). Indicators of Heart Disease (2024 UPDATE). Available online: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Sivarasan, R., & Mythili, M. S. (2023). Prediction System for Covid-19 Upcoming Cases Using Ensemble Classification. Journal of Advanced Applied Scientific Research, 5(4), 82–98. https://doi.org/10.46947/joaasr542023683

Rehman, A., Alam, T., Mujahid, M., Alamri, F. S., Al Ghofaily, B., & Saba, T. (2023). RDET stacking classifier: a novel machine learning based approach for stroke prediction using imbalance data. PeerJ Computer Science, 9. https://doi.org/10.7717/peerj-cs.1684

Peerbasha, S., Iqbal, Y. M., Praveen, K. P., Surputheen, M. M., & Saleem Raja, A. (2023). Diabetes Prediction using Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic Regression Classifiers. *Journal of Advanced Applied Scientific Research, 5*(4), 42–54. https://doi.org/10.46947/joaasr542023680

Statlog (Heart). UCI Machine Learning Repository. https://doi.org/10.24432/C57303. Accessed on 30-May-2024.

Wang, Y.-C., & Cheng, C.-H. (2021). A multiple combined method for rebalancing medical data with class imbalances. Computers in Biology and Medicine, 134, 104527. https://doi.org/10.1016/j.compbiomed.2021.104527

Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N., & Han, X. (2021). A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data. Information Sciences, 572, 574-589. https://doi.org/10.1016/j.ins.2021.02.056

Yewale, D., Vijayaragavan, S. P., & Bairagi, V. K. (2023). An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning. International Journal of Advanced Computer Science and Applications, 14(2). https://doi.org/10.14569/IJACSA.2023.0140223

Zhang, Y., Wang, G., Huang, X., & Ding, W. (2023). TSK fuzzy system fusion at sensitivity-ensemble-level for imbalanced data classification. Information Fusion, 92, 350-362. https://doi.org/10.1016/j.inffus.2022.12.014