Metaheuristic-Optimized Ensemble Learning for Glioma Grading

Lucas Silva Santana, Saulo Moraes Villela(✉), Luiz Maurílio da Silva Maciel, Raul Fonseca Neto, and Marcelo Bernardes Vieira

Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, Brazil
lucas.santana@estudante.ufjf.br, {saulo.moraes,luiz.maciel,raulfonseca.neto,marcelo.bernardes}@ufjf.br

Abstract. Accurate grading of gliomas is essential for clinical decision-making, prognosis estimation, and treatment planning. Recent advances in molecular profiling have improved diagnostic accuracy, but their routine use remains limited due to cost, availability, and annotation constraints in public datasets. Consequently, machine learning approaches based on clinical and mutation features have gained increasing attention. However, no single classifier consistently achieves optimal performance across heterogeneous biomedical datasets. In this work, we propose a metaheuristic-optimized ensemble learning framework for glioma grading that integrates multiple classical machine learning models using weighted soft voting. Particle Swarm Optimization is employed to automatically determine the optimal contribution of each classifier in the ensemble, targeting improvements in accuracy, recall-oriented F2-score, and area under the ROC curve. In particular, emphasizing recall-oriented metrics is crucial in clinical settings, as false negatives, misclassifying high-grade gliomas as low-grade, may lead to delayed diagnosis and inadequate treatment, potentially resulting in worse patient outcomes. The framework is evaluated on two public datasets derived from The Cancer Genome Atlas and the Chinese Glioma Genome Atlas using a rigorous nested cross-validation protocol. On TCGA, the proposed method achieves an accuracy of 87.8% and an F2-score of 90.1%. On the CGGA dataset, the proposed approach attains an F2-score of 92.5% and an accuracy of 76.2%. Overall, the findings indicate that metaheuristic-driven ensemble optimization provides a flexible strategy for glioma grading, allowing the model to be adapted to different evaluation objectives, particularly in clinical scenarios where sensitivity to high-grade tumors is essential.

Keywords: glioma grading · brain tumor · clinical and molecular features · ensemble learning · metaheuristic optimization

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 193–208, 2027.
https://doi.org/10.1007/978-3-032-30494-0_13

1 Introduction

Gliomas are the most common primary tumors of the central nervous system and represent a major challenge in neuro-oncology due to their biological heterogeneity and variable clinical behavior [3]. Accurate tumor classification is a critical step in clinical decision-making, directly influencing treatment planning, therapeutic response assessment, and estimation of prognosis [16]. Gliomas are broadly categorized into low-grade and high-grade tumors, with glioblastoma multiforme (GBM) being the most aggressive form. Despite advances in diagnostic protocols, reliable classification remains a complex task due to overlapping histopathological features and variability between observers [1–3].

Recent updates in the World Health Organization (WHO) classification have incorporated molecular alterations, such as isocitrate dehydrogenase (IDH) mutation status, to improve diagnostic accuracy. Although these molecular markers provide significant clinical value, their routine use is constrained by high costs, long turnaround times, and limited availability of comprehensive clinical annotation in publicly accessible datasets. These challenges motivate the development of computational methods capable of leveraging existing clinical and molecular data more effectively while minimizing the use of extensive molecular testing [10,12].

Machine learning (ML) has demonstrated strong potential in medical decision-support systems by enabling data-driven, objective, and reproducible analyses. Classical ML models, including Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (KNN), and gradient boosting-based methods such as Light Gradient Boosting Machine (LightGBM), have been widely applied to biomedical classification tasks due to their robustness and effectiveness in handling high-dimensional tabular data [9]. However, no single ML algorithm consistently achieves optimal performance across all datasets and problem settings, as formalized by the no-free-lunch theorem [15].

To address this limitation, ensemble learning approaches have been proposed to combine multiple classifiers, taking advantage of their complementary strengths to improve predictive performance and generalization [5,13]. Although conventional ensemble strategies such as bagging, boosting, and voting have achieved notable success, their effectiveness often depends on manually defined configurations and fixed combination schemes. Consequently, there is growing interest in adaptive and optimization-driven frameworks that can automatically determine optimal model combinations and configurations.

Metaheuristic optimization algorithms provide a powerful mechanism for navigating complex and high-dimensional search spaces and have been successfully applied to feature selection, hyperparameter tuning, and model combination tasks [6,8]. By integrating metaheuristics with classical ML models, it becomes possible to construct flexible ensemble systems that dynamically optimize classifier selection and combination strategies, leading to improved performance and robustness in heterogeneous datasets.

In this work, we propose a metaheuristic-based ensemble framework that integrates multiple classical machine learning models for glioma grading using

clinical and molecular characteristics. We employ Particle Swarm Optimization (PSO) [4] to optimize the weighting of individual classifiers within the ensemble, enabling objective-driven optimization tailored to specific evaluation criteria. In particular, the proposed framework allows prioritizing recall-oriented metrics, such as the F2-score, which are especially relevant in healthcare and clinical decision-making scenarios where false negatives may lead to delayed or inadequate treatment. In addition to recall-focused optimization, the framework can also target Accuracy and Area Under the Curve (AUC), providing flexibility to adapt the ensemble behavior to different clinical and methodological requirements. The PSO implementation used in this study is based on the Mealpy optimization library [14].

2 Related Work

Machine learning (ML) approaches for glioma grading and classification have been studied due to their potential to support clinical decision-making. Recent studies have explored a wide range of classifiers, feature selection strategies, and ensemble learning paradigms to address the heterogeneity and complexity of glioma grading.

Tasci et al. [12] proposed a hierarchical voting-based framework for glioma grading that integrates clinical and molecular features from The Cancer Genome Atlas (TCGA) and the Chinese Glioma Genome Atlas (CGGA) datasets. Their methodology employs multiple feature selection techniques, including Weight of Evidence, Recursive Feature Elimination, Random Forest-based importance, and Least Absolute Shrinkage and Selection Operator (LASSO), whose outputs are combined through a hierarchical voting scheme to identify robust and discriminative feature subsets. The selected features are subsequently evaluated using several supervised classifiers, namely Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Random Forest, and AdaBoost. Experimental results demonstrated that the proposed hierarchical voting-based feature selection strategy yields improved performance compared to the use of LASSO alone. They achieved an accuracy of 87.606% in the TCGA dataset and 79.668% in the CGGA dataset.

In a subsequent work, Tasci et al. [10] proposed the GradWise framework, which focuses on improving glioma grading through a hybrid feature selection strategy rather than ensemble classification. Their approach combines an embedded method (LASSO) with a filter-based method (mRMR) in a rank-based weighting scheme to identify the most relevant clinical and molecular features. They evaluated the selected feature subsets using five supervised models, including Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Random Forest, and AdaBoost. The experimental results showed that the proposed method achieved an accuracy of 87.007% in the TCGA dataset and 80.412% in the CGGA dataset.

Noviandy et al. [7] introduced a Light Gradient Boosting Machine (LightGBM) model integrated with Explainable Artificial Intelligence (XAI) techniques, specifically SHapley Additive exPlanations (SHAP), for glioma grading.

Their approach achieved a classification accuracy of 89.88% in the TCGA dataset by identifying key molecular features that strongly influence model predictions, thereby improving interpretability. However, their experiments were limited to the TCGA dataset, and no evaluation was conducted on the CGGA dataset, which restricted the assessment of the generalizability of the model among different populations. Furthermore, although the authors evaluated six different machine learning models for performance comparison, hyperparameter tuning was conducted exclusively for the LightGBM classifier, and they did not investigate ensemble learning or classifier combination strategies. It is also important to note that the experimental evaluation was performed using a single hold-out split, with 80% of the data used for training and 20% for testing, without employing k-fold cross-validation. As a result, the reported performance may be biased by the particular data partition used, making the results less robust and not directly comparable to approaches evaluated under cross-validation protocols, such as the method proposed in the present work.

As shown in Table 1, existing studies rely on fixed configurations or focus on feature selection without optimizing model combination or prioritizing clinically relevant metrics such as recall. To address these limitations, this work proposes a PSO-optimized ensemble that dynamically adjusts classifier contributions according to specific evaluation objectives, particularly in clinical scenarios where sensitivity to high-grade tumors is essential.

Table 1. Summary of related work.

Reference	Approach	Dataset(s)	Limitations
Tasci et al. [12]	Hierarchical voting-based ensemble combining multiple feature selection methods and classifiers	TCGA, CGGA	Lacks adaptive, objective-driven optimization mechanisms to adjust classifier contributions according to clinically relevant evaluation metrics, such as recall-oriented measures.
Tasci et al. [10]	Hybrid feature selection with rank-based weighting	TCGA, CGGA	Focuses on feature selection and does not explore ensemble learning optimization of multiple models
Noviandy et al. [7]	LightGBM model with explainability using SHAP	TCGA	Limited to a single dataset; lacks robust validation (no k-fold cross-validation); does not use ensemble methods or model combination optimization

3 Methodology

This section describes the data pre-processing steps, the ML models employed, the hyperparameter optimization strategy, the ensemble model combination approach based on metaheuristic optimization, and the evaluation metrics used in this study.

3.1 Datasets

In this work, we used two publicly available glioma datasets: the Glioma Grading Clinical and Mutation Features dataset derived from TCGA [11] and the dataset derived from CGGA [17].

The TCGA-based dataset was constructed to investigate the relationship between clinical characteristics, gene mutations, and glioma grading, specifically distinguishing between lower-grade gliomas (LGG) and glioblastoma multiforme (GBM). It consists of 839 samples described by 23 features, including 20 molecular and 3 clinical variables. The molecular features are binary, indicating the presence or absence of mutations in specific genes, and are encoded as 0 (not mutated) or 1 (mutated). Clinical features include gender, age, and race. Gender is encoded as 0 for male and 1 for female. Race is encoded as 0 for White, 1 for Black or African American, 2 for Asian, and 3 for American Indian or Alaska Native.

The molecular features of the TCGA dataset comprise the following genes: IDH1, TP53, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, GRIN2A, IDH2, FAT4, and PDGFRA.

In addition, a second dataset derived from the CGGA was used. The CGGA dataset contains 286 samples described by 22 features. In contrast to the TCGA dataset, the CGGA dataset does not include the race feature, as all samples correspond to individuals from the Chinese population. Apart from this difference, the feature definitions and encoding schemes are consistent with those of the TCGA dataset.

For both datasets, glioblastoma multiforme (GBM) was defined as the positive class (label 1), while lower-grade glioma (LGG) was defined as the negative class (label 0) for all performance metrics evaluated.

3.2 Pre-processing

Prior to model training and evaluation, a pre-processing step was applied to ensure numerical stability and feature comparability. The numerical feature age was normalized using Min–Max normalization, scaling its values to the $[0, 1]$ interval. To prevent data leakage, the normalization parameters were computed exclusively on the training data within each outer cross-validation fold and subsequently applied to the corresponding test fold. This procedure is particularly important for distance-based and margin-based classifiers, such as k-Nearest Neighbors (KNN) and Support Vector Machines (SVM), which are sensitive to

feature scaling. In the original dataset, the categorical variable race is encoded as an integer-valued feature. However, in our pre-processing pipeline, this variable is treated as categorical and transformed using one-hot encoding to avoid introducing artificial ordinal relationships between categories. All remaining features are binary and were therefore not subjected to additional scaling or transformation. In this study, we did not apply data balancing or resampling techniques.

3.3 Machine Learning Models

We conducted experiments using a diverse set of classical supervised machine learning models that are widely adopted in biomedical classification tasks. The evaluated models include Logistic Regression, Support Vector Machine with RBF kernel (SVM RBF), k-Nearest Neighbors (KNN), Decision Tree, Random Forest, AdaBoost, Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), and Naive Bayes. This selection enables the evaluation of both linear and nonlinear decision boundaries, as well as tree-based ensemble methods, boosting techniques, and probabilistic classifiers, providing complementary perspectives on the underlying data distribution.

3.4 Hyperparameter Tuning

To identify the optimal hyperparameter configurations for each model, we employed a random search strategy. The random search was selected due to its efficiency in exploring high-dimensional hyperparameter spaces compared to exhaustive grid search. The hyperparameter search space for each model is detailed in Table 2.

We evaluated the dataset using a nested cross-validation strategy. The outer loop consisted of a 10-fold cross-validation used to estimate the generalization performance, where, at each iteration, the data were split into training and test subsets. Within each training subset, we performed an inner 5-fold cross-validation for model selection and hyperparameter tuning, using the F2-score as the optimization criterion. In this inner loop, the training data were further split into training and validation sets, and the validation set was used exclusively for hyperparameter optimization.

The optimized model configurations were then evaluated on the corresponding test fold. This nested evaluation strategy ensures that each sample in the dataset is used for testing exactly once, reducing dependency on a single train-test split and mitigating the risk of overfitting. Consequently, this approach provides a more reliable and unbiased estimate of the model's performance.

3.5 Model Combination Strategy

To further improve classification performance, we employed ensemble learning strategies to combine the predictions of multiple base classifiers. Initially, both soft voting and hard voting ensembles were evaluated. In soft voting, the predicted class probabilities of the individual classifiers are averaged with equal

Table 2. Hyperparameter search space explored for each model using random search, specifying value ranges, discretization steps, and sampling strategies (uniform and logarithmic scales).

Algorithm	Hyperparameter	Search space
AdaBoost	Number of estimators	[50, 500] (step = 50)
	Learning rate	[0.01, 1.0] (uniformly spaced, 5 values)
Decision Tree	Criterion	{gini, entropy}
	Maximum depth	[10, 30] (step = 10)
	Maximum features	{sqrt, log2}
KNN	Number of neighbors	[5, 25] (step = 5)
	Weights	{uniform, distance}
	Algorithm	{auto, brute}
LightGBM	Number of estimators	[50, 500] (step = 50)
	Minimum child samples	[3, 15] (step = 3)
Logistic Regression	Penalty	{l1, l2}
	C	[0.1, 100] (uniformly spaced, 25 values)
Naive Bayes	Var smoothing	$[10^{-12}, 10^{-3}]$
Random Forest	Number of estimators	[50, 500] (step = 50)
	Maximum features	{sqrt, log2}
SVM (RBF)	C	$[10^{-3}, 10^{3}]$ (log-scale, 10 values)
	Gamma	$[10^{-3}, 10^{1}]$ (log-scale, 5 values)
XGBoost	Number of estimators	[50, 500] (step = 50)
	Learning rate	[0.01, 1.0] (uniformly spaced, 5 values)

weights, allowing the final decision to incorporate model confidence. In contrast, hard voting determines the final prediction based on the majority of discrete class labels produced by the individual models.

Since individual classifiers may exhibit different levels of reliability, a weighted soft voting ensemble was also investigated. In this approach, a weight is assigned to each classifier, and the final prediction is obtained as a weighted combination of the predicted class probabilities, i.e., $\sum_{i=1}^{M} w_i \cdot p_i$, where p_i denotes the predicted probability of the positive class produced by the i-th classifier. The weights are constrained such that $\sum_{i=1}^{M} w_i = 1$. Algorithm 1 describes the training and model combination process.

The optimal set of weights was determined using the PSO metaheuristic, with the objective of maximizing specific performance metrics, namely Accuracy (ES ACC), F2 (ES F2) and AUC (ES AUC). PSO is a population-based optimization algorithm in which each particle represents a candidate solution that iteratively moves through the search space according to its own experience and that of neighboring particles.

Algorithm 1. Training and Evaluation with Nested Cross-Validation

Input: Dataset (X, y), set of base models $\mathcal{M}$
Output: Predictions for individual and ensemble models
Define 10-fold outer cross-validation
Define 5-fold inner cross-validation
for each outer fold **do**
  for each base model $m \in \mathcal{M}$ **do**
    Perform Randomized Search with inner CV optimizing F2-score
    Select the best model configuration
    Generate and store predictions on the outer test set
  end for
  Apply soft and hard voting ensembles and store outer test set predictions
  Optimize ensemble weights using PSO on the outer training data
  Apply optimized weights to the outer test set predictions and store results
end for
Compute evaluation metrics using aggregated predictions across all outer folds

At each iteration, the velocity of a particle is updated according to:

$$\mathbf{v}_i^{t+1} = w\mathbf{v}_i^t + c_1 r_1(\mathbf{p}_i - \mathbf{x}_i^t) + c_2 r_2(\mathbf{g} - \mathbf{x}_i^t), \tag{1}$$

where $\mathbf{v}_i^t$ and $\mathbf{x}_i^t$ denote the velocity and position of particle i at iteration t, respectively; $\mathbf{p}_i$ is the personal best position found by the particle; $\mathbf{g}$ is the global best position found by the swarm; w is the inertia weight; c_1 and c_2 are the cognitive and social coefficients; and r_1, r_2 are random numbers uniformly distributed in $[0, 1]$.

The particle position is then updated as:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1}. \tag{2}$$

The default parameter settings of the PSO implementation were used, namely epoch = 1000, population size = 50, $c_1 = 2.05$, $c_2 = 2.05$, and $w = 0.4$. These parameters control the balance between exploration and exploitation during the optimization process.

In our approach, each particle in the PSO population represents a candidate solution encoding the weights assigned to the base classifiers in the weighted ensemble. A particle is defined as a real-valued vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)$, where each component w_i corresponds to the weight associated with the i-th classifier, and M denotes the total number of base models. Therefore, the dimensionality of each particle is equal to the number of classifiers in the ensemble.

The optimization process was conducted exclusively on the training data within the nested cross-validation framework to avoid information leakage. Once the optimal weights were identified, they were fixed and applied to the independent test data, ensuring a fair and unbiased evaluation of the generalization performance of the weighted ensemble.

3.6 Evaluation Metrics

The performance of our model was assessed using multiple evaluation metrics to provide an analysis of its classification effectiveness. These metrics include accuracy, precision, recall, F1-score, F2-score, and area under the receiver operating characteristic curve (AUC). Although accuracy provides an overall measure of correctness, precision and recall capture complementary aspects of classification performance, particularly in the presence of class imbalance. The F1-score balances precision and recall equally, whereas the F2-score emphasizes recall, which is often critical in medical decision-making contexts. The AUC metric evaluates the model's discriminative ability across varying classification thresholds.

3.7 Experimental Setup

All experiments were conducted on a computer running Linux Mint 21.3, equipped with an Intel Core i7-12700 CPU (12th generation) operating at 2.10 GHz and 32 GB of RAM. We implemented all experiments in Python version 3.11, using scikit-learn version 1.7.2, LightGBM version 4.6.0, and the mealpy optimization library version 3.0.3. To ensure reproducibility, we fixed the random seed to 1 across all experiments.

4 Results

Initially, we conducted experiments using individual machine learning classifiers trained independently, without any ensemble or voting strategy. For each classifier, we performed hyperparameter optimization using the nested cross-validation approach described previously. Table 3 presents the results of the individual models evaluated on the TCGA dataset in the present study.

Table 3. Results of Individual Models on the TCGA dataset.

Model	ACC	PRE	REC	F1	F2	AUC
AdaBoost	86.65	79.27	**92.33**	85.30	**89.38**	90.20
Decision Tree	85.82	78.48	91.19	84.36	88.33	89.72
KNN	86.17	78.64	92.05	84.82	89.01	**90.99**
LightGBM	86.29	78.83	92.05	84.93	89.06	90.09
Logistic Regression	**87.60**	**82.12**	90.06	**85.91**	88.35	90.68
Naive Bayes	86.17	80.26	88.92	84.37	87.04	88.33
Random Forest	87.13	81.28	90.06	85.44	88.15	89.98
SVM RBF	86.65	79.41	92.05	85.26	89.21	90.19
XGBoost	85.46	78.33	90.34	83.91	87.65	90.25

All evaluated models achieved consistent performance, with accuracy values above 85% and AUC scores close to 0.90, indicating good overall discrimination capability.

In addition to accuracy and AUC, recall and F2 metrics were emphasized due to their relevance in clinical decision-making. AdaBoost, KNN, LightGBM, and SVM with the RBF kernel achieved recall values above 92%, which are reflected in their higher F2-scores. These results indicate a good ability to identify positive cases while prioritizing sensitivity. Although Logistic Regression and Random Forest achieved higher accuracy and precision, their recall and F2-score values remained competitive, indicating a balanced performance between precision and recall.

Table 4 presents the performance of the individual classifiers on the CGGA dataset.

Table 4. Results of Individual Models on the CGGA dataset.

Model	ACC	PRE	REC	F1	F2	AUC
AdaBoost	65.73	71.11	82.90	76.56	80.24	59.49
Decision Tree	65.73	69.71	87.05	77.42	82.92	65.24
KNN	71.68	75.93	84.97	80.20	83.00	71.52
LightGBM	71.33	74.89	86.53	80.29	83.92	71.19
Logistic Regression	70.63	72.80	**90.16**	**80.56**	**86.05**	65.17
Naive Bayes	**72.38**	**78.50**	81.35	79.90	80.76	**74.66**
Random Forest	68.88	77.08	76.68	76.88	76.76	71.07
SVM RBF	68.88	73.21	84.97	78.66	82.33	61.78
XGBoost	67.13	72.20	83.42	77.40	80.90	66.93

Overall, the results indicate moderate classification performance, with accuracy values ranging from 65.73% to 72.38% and AUC scores varying between 59.49 and 74.66, suggesting a more challenging classification scenario compared to the TCGA dataset.

With respect to recall-focused metrics, Logistic Regression achieved the highest recall (90.16%) and the best F1 and F2-scores, indicating a strong sensitivity to positive cases. Decision Tree and LightGBM also exhibited high recall values above 86%, which is reflected in their competitive F2-scores. Naive Bayes obtained the highest accuracy (72.38%), precision (78.50%), and AUC (74.66).

Overall, the results show that different models emphasize different performance aspects on the CGGA and TCGA datasets. These findings motivate the evaluation of ensemble strategies to combine complementary model behaviors and improve overall performance.

Table 5 presents the results of the ensemble strategies evaluated on the TCGA dataset, including soft voting, hard voting, and weighted ensemble models optimized using PSO. AUC was not computed for the hard voting scheme, as it produces only discrete class labels rather than continuous confidence scores required

for ROC analysis. We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. We constructed the ensemble combinations using three different sets of base classifiers derived from Table 3: all evaluated models, a subset of five models (AdaBoost, KNN, LightGBM, Logistic Regression, and Naive Bayes), and a subset of four models (AdaBoost, KNN, Logistic Regression, and Naive Bayes).

Table 5. Ensemble Learning Results on the TCGA dataset.

Models	Method	ACC	PRE	REC	F1	F2	AUC
All	Soft Voting	87.01	80.60	90.91	85.45	88.64	91.31
	Hard Voting	87.01	80.60	90.91	85.45	88.64	–
	ES ACC	87.60	81.16	91.76	86.13	89.42	91.10
	ES F2	87.37	80.45	92.33	85.98	89.68	91.27
	ES AUC	87.13	80.20	92.05	85.71	89.40	91.28
5	Soft Voting	87.13	80.81	90.91	85.56	88.69	91.04
	Hard Voting	86.89	80.40	90.91	85.33	88.59	–
	ES ACC	**87.84**	**81.25**	92.33	**86.44**	89.88	91.30
	ES F2	86.89	79.95	91.76	85.45	89.13	90.82
	ES AUC	87.25	80.40	92.05	85.83	89.45	91.31
4	Soft Voting	87.13	80.81	90.91	85.56	88.69	91.09
	Hard Voting	86.77	80.66	90.06	85.10	88.01	–
	ES ACC	87.60	81.16	91.76	86.13	89.42	91.51
	ES F2	87.60	80.54	**92.90**	86.28	**90.13**	91.39
	ES AUC	87.60	80.69	92.61	86.24	89.96	**91.55**

The ensemble-based approaches exhibit performance comparable to, and in most cases better than, the individual classifiers.

Both hard and soft voting achieved comparable results in terms of accuracy, precision, recall, and F2-score, while soft voting achieved a higher AUC (91.31) than the best individual classifier (90.99%), indicating improved class separation. The PSO-based ensembles further enhanced performance by optimizing classifier weights according to different objective functions. The ensemble optimized for accuracy achieved the highest overall accuracy (87.84%) and F1-score (86.44%), while the ensemble optimized for F2 attained the highest recall (92.90%) and F2-score (89.68%), reinforcing its effectiveness in recall-sensitive scenarios. The ensemble optimized for AUC achieved the highest AUC among all evaluated combinations (91.55), reflecting improved class discrimination across decision thresholds. This performance surpasses the best individual classifier (AUC of 90.99), confirming the benefit of the ensemble approach for enhancing discriminative capability.

Overall, these results highlight the flexibility of the PSO-based ensemble framework, which allows for prioritizing different performance criteria while maintaining balanced and competitive classification performance.

Table 6 reports the performance of ensemble strategies on the CGGA dataset, including soft voting, hard voting, and PSO-based weighted ensembles optimized for accuracy, F2-score, and AUC. We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. Compared to the TCGA dataset, the overall performance on the CGGA dataset is lower across all metrics, confirming the increased difficulty of this classification task.

Table 6. Ensemble Learning Results on the CGGA dataset.

Models	Method	ACC	PRE	REC	F1	F2	AUC
All	Soft Voting	73.43	74.89	91.19	82.24	87.39	73.55
	Hard Voting	71.68	74.56	88.08	80.76	85.00	–
	ES ACC	73.08	75.89	88.08	81.53	85.34	73.02
	ES F2	**76.22**	74.51	**98.45**	**84.82**	**92.50**	73.97
	ES AUC	72.03	75.34	87.05	80.77	84.42	69.30
5	Soft Voting	73.78	74.58	92.75	82.68	88.44	**75.67**
	Hard Voting	74.83	74.69	94.82	83.56	89.97	–
	ES ACC	72.38	76.64	84.97	80.59	83.16	72.98
	ES F2	71.33	71.60	95.34	81.78	89.41	72.88
	ES AUC	70.98	**77.50**	80.31	78.88	79.73	74.15
4	Soft Voting	74.13	73.71	95.85	83.33	90.42	74.87
	Hard Voting	73.78	75.00	91.71	82.52	87.80	–
	ES ACC	71.68	76.17	84.46	80.10	82.66	72.07
	ES F2	69.23	70.27	94.30	80.53	88.26	71.72
	ES AUC	71.33	76.81	82.38	79.50	81.21	74.07

Among the PSO-based ensembles, the model optimized for F2-score obtained the best overall results, achieving the highest accuracy (76.22%), recall (98.45%), F1-score (84.82%), and F2-score (92.50%), emphasizing its effectiveness in recall-oriented optimization. These results indicate that recall-oriented optimization effectively emphasizes the correct identification of high-grade gliomas, which is particularly important in clinical decision-making scenarios where false negatives may lead to delayed or inadequate treatment.

In contrast, the ensemble optimized for accuracy did not outperform the voting-based methods on CGGA. When compared to the TCGA results, the CGGA ensembles exhibit smaller absolute gains over voting strategies, indicating that the benefits of weighting classifiers are more dataset-dependent and more pronounced in the TCGA scenario.

Table 7 presents the results of the best-performing ensemble optimized for the F2-score on the TCGA dataset and compares it with state-of-the-art methods.

We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. The proposed approach employs PSO to determine optimal weights for each classifier, with a particular focus on improving the identification of positive cases when optimizing for the F2-score, while maintaining competitive overall performance. In addition, the table also reports the best results obtained by ensembles optimized for accuracy and AUC, illustrating the flexibility of the proposed framework in addressing different performance objectives.

Table 7. Comparison of our approach with state-of-the-art methods on the TCGA dataset.

Method	ACC	PRE	REC	F1	F2	AUC
Tasci et al. [12]	87.6	**85.8**	81.5	85.8	82.3	91.4
Tasci et al. [10]	87.0	80.4	91.5	85.5	89.0	91.1
ES ACC	**87.8**	81.3	92.3	**86.4**	89.9	91.3
ES F2	87.6	80.5	**92.9**	86.3	**90.1**	91.4
ES AUC	87.6	80.7	92.6	86.2	90.0	**91.6**

The main objective of our method is to prioritize the correct classification of the positive class while preserving competitive accuracy. When compared to the hierarchical approach proposed in Tasci et al. [12], the ensemble optimized for the F2-score achieves substantially higher recall and a higher F2-score for the positive class while maintaining equal accuracy. In comparison with the grad-wise method presented in Tasci et al. [10], the proposed ensemble optimized for the F2-score attains improved performance across all reported metrics. Moreover, the ensemble optimized for AUC outperforms both state-of-the-art methods, demonstrating superior overall discriminative capability. Finally, the ensemble optimized for accuracy achieves a higher accuracy than both reference methods, while also yielding higher recall, F1-score, and F2-score, further highlighting the effectiveness and flexibility of the proposed ensemble framework.

Overall, the higher recall, F2-score and AUC obtained by the proposed method demonstrate that the PSO-based weighting strategy effectively emphasizes the positive class, enabling improved sensitivity while maintaining competitive performance on the TCGA dataset.

Table 8 presents the results the best-performing ensembles on the CGGA dataset, optimized with respect to the F2-score, and compares it with state-of-the-art methods. We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. The table also reports the best results obtained by ensembles optimized for accuracy and AUC. The experimental results indicate that the CGGA dataset represents a more challenging classification problem, as reflected by the lower accuracy.

Despite the lower accuracy, the proposed method demonstrates substantially superior performance in recall-oriented metrics, such as F2-score. Our method

Table 8. Comparison of our approach with state-of-the-art methods on the CGGA dataset.

Method	ACC	PRE	REC	F1	F2	AUC
Tasci et al. [12]	79.7	76.4	59.3	64.8	62.1	78.8
Tasci et al. [10]	**80.4**	**80.7**	60.7	67.9	63.9	**79.8**
ES ACC	73.1	75.9	88.1	81.5	85.3	73.0
ES F2	76.2	74.5	**98.5**	**84.8**	**92.5**	74.0
ES AUC	71.0	77.5	80.3	78.9	79.7	74.2

explicitly allows optimizing the ensemble toward different clinical priorities. In particular, the optimization of the F2-score emphasizes recall, which is highly relevant in clinical decision-making for glioma grading. In this context, false negatives (i.e., misclassifying high-grade gliomas as low-grade) may lead to delayed or inadequate treatment, potentially resulting in significantly worse patient outcomes. Therefore, improving recall from approximately 60% to up to 98.5% on CGGA represents a substantial clinical gain, even at the expense of moderate reductions in accuracy and AUC. These results highlight the effectiveness of the proposed PSO-based ensemble in emphasizing sensitivity to the positive class.

5 Conclusion

In this work, we proposed a metaheuristic-based ensemble learning framework for glioma grading using clinical and molecular features. By optimizing ensemble weights with Particle Swarm Optimization (PSO), the proposed method enables objective-driven optimization tailored to different evaluation criteria, including Accuracy, F2-score, and AUC.

Results on the TCGA and CGGA datasets showed that the proposed ensemble outperforms individual classifiers and conventional voting strategies, particularly when optimized for the F2-score, achieving higher recall and F2-score values while maintaining competitive accuracy. Although the CGGA dataset presents a more challenging classification task due to its reduced sample size, the proposed method preserved robust recall-oriented performance across both cohorts.

Overall, these findings suggest that metaheuristic-optimized ensembles offer a practical and adaptable solution for glioma grading using structured clinical and molecular features. Although evaluated on glioma grading, the proposed framework is applicable to other recall-sensitive classification tasks. Future work can explore larger and more feature-rich datasets to further evaluate the robustness and generalization of the proposed approach.

Acknowledgments. The authors would like to thank the Minas Gerais State Research Foundation - FAPEMIG.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Eder, K., Kalman, B.: Molecular heterogeneity of glioblastoma and its clinical relevance. Pathol. Oncol. Res. **20**(4), 777–787 (2014). https://doi.org/10.1007/s12253-014-9833-3
2. Hanif, F., Muzaffar, K., Perveen, K., Malhi, S.M., Simjee, S.U.: Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment. Asian Pac. J. Cancer Prev. APJCP 18(1), 3 (2017). https://doi.org/10.22034/APJCP.2017.18.1.3
3. Horbinski, C., Berger, T., Packer, R.J., Wen, P.Y.: Clinical implications of the 2021 edition of the WHO classification of central nervous system tumours. Nat. Rev. Neurol. **18**(9), 515–529 (2022). https://doi.org/10.1038/s41582-022-00679-w
4. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-International Conference on Neural Networks. vol. 4, pp. 1942–1948. IEEE (1995). https://doi.org/10.1109/ICNN.1995.488968
5. Mahajan, P., Uddin, S., Hajati, F., Moni, M.A.: Ensemble learning for disease prediction: a review. Healthcare **11**(12), 1808 (2023). https://doi.org/10.3390/healthcare11121808
6. Nematzadeh, S., Kiani, F., Torkamanian-Afshar, M., Aydin, N.: Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: a bioinformatics study on biomedical and biological cases. Comput. Biol. Chem. **97**, 107619 (2022). https://doi.org/10.1016/j.compbiolchem.2021.107619
7. Noviandy, T.R., Idroes, G.M., Hardi, I.: Integrating explainable artificial intelligence and light gradient boosting machine for glioma grading. Inf. Health **2**(1), 1–8 (2025). https://doi.org/10.1016/j.infoh.2024.12.001
8. Nssibi, M., Manita, G., Korbaa, O.: Advances in nature-inspired metaheuristic optimization for feature selection problem: a comprehensive survey. Comput. Sci. Rev. **49**, 100559 (2023). https://doi.org/10.1016/j.cosrev.2023.100559
9. Reel, P.S., Reel, S., Pearson, E., Trucco, E., Jefferson, E.: Using machine learning approaches for multi-omics data analysis: a review. Biotechnol. Adv. **49**, 107739 (2021). https://doi.org/10.1016/j.biotechadv.2021.107739
10. Tasci, E., Jagasia, S., Zhuge, Y., Camphausen, K., Krauze, A.V.: GradWise: a novel application of a rank-based weighted hybrid filter and embedded feature selection method for glioma grading with clinical and molecular characteristics. Cancers **15**(18), 4628 (2023). https://doi.org/10.3390/cancers15184628
11. Tasci, E., Zhuge, Y., Camphausen, K., Krauze, A.: Glioma grading clinical and mutation features. UCI Mach. Learn. Repository (2022). https://doi.org/10.24432/C5R62J
12. Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., Krauze, A.V.: Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. Int. J. Mol. Sci. **23**(22), 14155 (2022). https://doi.org/10.3390/ijms232214155
13. Tiwari, A., Chugh, A., Sharma, A.: Ensemble framework for cardiovascular disease prediction. Comput. Biol. Med. **146**, 105624 (2022). https://doi.org/10.1016/j.compbiomed.2022.105624
14. Van Thieu, N., Mirjalili, S.: MEALPY: an open-source library for latest meta-heuristic algorithms in python. J. Syst. Architect. **139**, 102871 (2023). https://doi.org/10.1016/j.sysarc.2023.102871
15. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997). https://doi.org/10.1109/4235.585893

16. Zhan, Y., Hao, Y., Wang, X., Guo, D.: Advances of artificial intelligence in clinical application and scientific research of neuro-oncology: Current knowledge and future perspectives. Crit. Rev. Oncol. Hematol. 104682 (2025). https://doi.org/10.1016/j.critrevonc.2025.104682
17. Zhao, Z., et al.: Chinese glioma genome atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. Genomics Proteomics Bioinform. **19**(1), 1–12 (2021). https://doi.org/10.1016/j.gpb.2020.10.005