Glioma Classification Using Harris Hawks-Driven Optimized Gradient Boosting Classifier Along with SHAP-Based Interpretability

SM Naim[1], Jun-Jiat Tiang[2]*, Abdullah-Al Nahid[3]*
Electronics and Communication Engineering Discipline, Khulna University, Khulna 9208, Bangladesh[1,3]
Centre for Wireless Technology-Centre of Excellence for Intelligent Network-Faculty of Artificial Intelligence and Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia[2]

***Abstract*—Gliomas are considered one of the most lethal and aggressive types of brain cancer, responsible for countless deaths worldwide. This study seeks to improve glioma classification using cutting-edge machine learning (ML) techniques to differentiate between glioma subtypes based on clinical and genomic data. The goal is to identify important biomarkers and features influencing glioma classification, with an emphasis on improving feature selection and model interpretability. For glioma classification, the Gradient Boosting Classifier (GBC) was employed. The Harris Hawks Optimization (HHO) algorithm was used for feature selection and hyperparameter fine-tuning to enhance the model's performance. Additionally, SHapley Additive exPlanations (SHAP) were applied to improve model interpretability and to better understand feature contributions.The Gradient Boosting (GB) method yielded the best performance among the selected models, achieving an accuracy of 88.40%, precision of 87.3%, recall of 88.48%, and an F1 score of 88.29%, with feature selection and hyperparameter tuning using the Harris Hawks Optimization. These results highlight the significance of hyperparameter tuning and feature selection in enhancing classification performance. Key features such as IDH1, Age at Diagnosis, and EGFR were identified as the most influential in distinguishing glioma subtypes. SHAP analysis further confirmed the importance of these features in the model.This study shows that the Gradient Boosting Classifier (GBC), optimized with Harris Hawks Optimization (HHO), significantly improves glioma classification, achieving a high F1 score. Key features like IDH1, Age at Diagnosis, and EGFR were identified, showcasing its potential for enhanced glioma diagnosis.**

***Keywords*—Glioma; gradient boosting; Harris Hawks Optimization (HHO); SHAP; feature selection; interpretability; TCGA; IDH1; EGFR**

I. Introduction

Brain and nervous system cancer is one of the most dangerous cancers in today's age. The death toll is rising day by day. According to research by the World Health Organization, there were approximately 251,000 deaths from brain and nervous system cancer in 2020, across both sexes, up to 85 years of age [1]. Among these, gliomas represent a diverse group of tumours originating from glial cells in the brain and pose significant diagnostic and therapeutic challenges in neuro-oncology [2]. Accurate grading of gliomas is critical, as it informs clinical decision-making and influences patient outcomes. However, conventional grading methods often lack precision and are resource-intensive. As illustrated in Fig. 1, the global burden motivates the need for reliable, explainable decision support.

Survival rates for glioma patients remain low, especially for high-grade variants such as glioblastoma multiforme (GBM), which alone accounts for more than 60 percent of adult primary brain tumours [6], [7]. The 2007 WHO classification introduced a system based on histological features, which was later enhanced by incorporating molecular markers such as IDH mutations [4]. In recent years, the convergence of neuroinformatics, computational neuroimaging, and machine learning has provided promising avenues for improving glioma classification. Machine learning algorithms can analyse complex, high-dimensional data from genomic and clinical sources to uncover patterns that traditional methods may overlook. One of the key challenges in glioma grading is the reliance on expensive molecular tests like IDH1/IDH2 mutation analysis, which can cost between USD 135–1800 and take up to two weeks to process [9], [10]. Additionally, variables such as age, sex, and clinical symptoms also influence tumour behaviour but are not always integrated effectively due to limited dataset annotations [11].

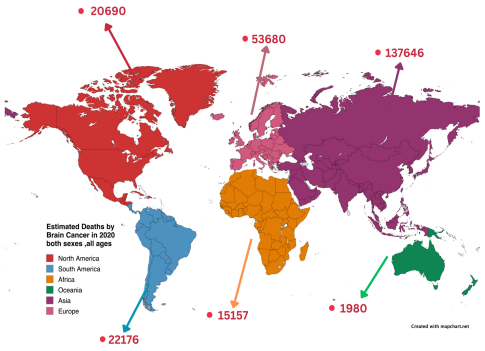


Fig. 1. Estimated number of brain cancer deaths by region in 2020.

Gliomas are among the most prevalent and deadly forms of brain cancer [3]. Early and accurate detection is crucial for prognosis and treatment decisions. Machine learning has shown great promise in extracting valuable insights from complex medical datasets, making it possible to predict disease

*Corresponding author.

outcomes with higher accuracy. This paper leverages advanced techniques such as Gradient Boosting Classifiers (GBC) and optimisation algorithms like HHO to identify key features that improve glioma diagnosis, with a specific focus on their interpretation through SHAP analysis [21].

Despite many promising studies, important gaps remain. A number of works maximize accuracy without explaining the predictions; others provide explanations but rely on only one data type (for example, imaging alone) or evaluate models using a single cross-validation loop that can overestimate performance. It is still uncommon to jointly perform feature selection and hyperparameter tuning on combined clinical+genomic predictors, then audit the final model with SHAP and report fold-aggregate results that better reflect generalization and clinical use.

This paper addresses these gaps with a simple, end-to-end pipeline that uses HHO with a GBC to jointly pick features and tune hyperparameters on TCGA-LGG/GBM data, followed by SHAP to check that the most influential variables—such as *IDH1*, *EGFR*, *TP53*, *NF1*, and Age—agree with contemporary glioma biology and the WHO-2021 system [5] [22]. We evaluate the model with cross-validation and summarize results across folds rather than relying on a single split. We also explain why gains in threshold-based metrics (accuracy, F1) can coexist with small changes in ROC AUC, and we outline how the outputs could support care by triaging cases for confirmatory molecular testing and by providing feature-level explanations suitable for tumor-board discussion.

II. Related Works

TCGA (The Cancer Genome Atlas) data have been used extensively to classify gliomas [8]. To differentiate glioblastoma multiforme (GBM) patients from others, Ko and Brody, for instance, used a gradient boosting classifier on TCGA copy-number data, obtaining an AUC of 0.875 [13]. Likewise, Sánchez-Marqués et al. demonstrated that ensemble approaches (e.g., CatBoost) outperformed standard classifiers on TCGA glioma cohorts by using TCGA-LGG and TCGA-GBM molecular and clinical data for glioma grade prediction [14]. These investigations emphasize how rich the TCGA-LGG/GBM dataset—which consists of hundreds of patients and clinical/molecular data—is for glioma subtype and grade classification problems. Underlining the relevance of TCGA in glioma research, other studies have built predictive models of glioma subtype or prognosis using TCGA-derived data (gene expression, mutations).

Medical and glioma classification has seen especially successful application for gradient boosting techniques. In oncology classification tasks, ensemble tree-boosting algorithms including XGBoost, LightGBM, and CatBoost routinely produce state-of-the-art results. Tang et al. used XGBoost on TCGA GBM transcriptomic profiles, for example, to classify GBM into its three subtypes (proneural, classical, mesenchymal), identifying a five-gene signature and approximating 80% accuracy [17]. Xia et al. trained GradientBoost and LightGBM models on MRI radiomic features to differentiate GBM from solitary metastases, achieving ROC-AUC > 0.90, and used SHAP to interpret feature contributions [16]. Gradient boosting approaches have also performed powerfully in radiomics-based brain tumor studies. Using MRI data, Kha et al. built an XGBoost model to forecast 1p/19q co-deletion in lower-grade glioma and used SHAP to choose the most useful radiomic features [15]. These and other studies—e.g., Sánchez-Marqués et al. using CatBoost [14]—show that gradient boosting classifiers are well-suited to complex, high-dimensional medical data, producing high accuracy in glioma grading and biomarker prediction.

Applied for feature selection and hyperparameter tuning in biomedical machine learning, Harris Hawks Optimization (HHO) is a new nature-inspired metaheuristic. Originally proposed in 2019 by Heidari et al., HHO has been modified for high-dimensional biomedical data in later work. Elgamal et al., for instance, chose ideal feature subsets in medical datasets by means of an enhanced HHO coupled with simulated annealing. Pirgazi et al. proposed a two-stage filter-wrapper approach whereby an enhanced HHO-based wrapper (with GRASP) further refines features chosen by a filter to identify an optimal subset for classification [21]. Apart from feature selection, HHO has been applied effectively to adjust model hyperparameters in clinical predictive models. Using HHO in a COVID-19 detection system, Kumar et al. optimized the hyperparameters of boosting classifiers (XGBoost, LightGBM, etc.), improving model performance and enabling integration of SHAP for feature analysis [18]. In a liver cirrhosis prediction task, Nalasari et al. similarly combined XGBoost with HHO for hyperparameter tuning, reporting notably better accuracy and less overfitting than standard XGBoost [19]. In the neuro-oncology setting [12], Kurdi et al. embedded HHO into a convolutional neural network (HHOCNN) for MRI-based brain tumor detection; their HHO-optimized CNN attained 98% accuracy on benchmark MRI datasets [20]. These cases show that by automatically selecting features or tuning parameters, HHO can efficiently improve model training; moreover, its application to biomedical classification is becoming rather common.

Rising in prominence in healthcare, interpretable artificial intelligence techniques such as SHapley Additive exPlanations (SHAP) offer insights into black-box model decisions. Several glioma research studies have used SHAP to clarify predictive characteristics. Before training the XGBoost model for 1p/19q status, Kha et al. ranked radiomic features using SHAP values, so increasing the transparency of the model [15]. Xia et al. calculated mean SHAP values for radiomic features in imaging studies to underline the most important predictors of tumor type [16]. Using SHAP to offer both global and local explanations of model outputs in a glioma survival-prediction web tool helped to show how clinical and molecular factors influence risk projections. Likewise, Kumar et al. used SHAP analysis on their COVID-19 boosting model to find important risk factors and simplify the predictions [18]. These papers demonstrate that SHAP is a useful tool in medical machine learning since it quantifies feature importance in complicated models, thus improving clinician confidence. SHAP has been applied not only for MRI/radiomics models but also for genomic and clinical models in the glioma domain, exposing which genetic changes or patient factors drive predictions.

All told, previous studies have shown the value of TCGA-based ML for glioma, the efficacy of gradient-boosting classifiers in such tasks, the promise of HHO for optimizing feature

sets and model parameters, and the importance of SHAP in model interpretation. These components have, nevertheless, mainly been studied separately. The present work validates the model with SHAP-based interpretability and combines these developments by using HHO to execute both feature selection and hyperparameter tuning of a gradient boosting classifier on TCGA glioma data. We know of no prior reports of this combined approach: an HHO-optimized GBC trained on TCGA genomic data using SHAP validation. It jointly improves model performance and transparency, bridging gaps in current work. While prior studies have shown that gradient boosting can achieve strong accuracy, many do not explain predictions in a way that is practical for clinical discussion, and others rely on a single data type rather than integrating clinical and genomic information. Several works also tune models within a single cross-validation loop, which can make the reported numbers look stronger than they are in routine use. As a result, three limitations remain common: modest gains in accuracy without clear interpretability, limited use of combined clinical+genomic inputs, and evaluation protocols that do not emphasize generalization. Our approach is designed to address these points by jointly selecting features and tuning hyperparameters with HHO, auditing the final model with SHAP, and reporting fold-level summaries that reflect how the model is likely to behave beyond one split.

III. Methodology

Based on the dataset, this work has carried out a classification task to distinguish between patients with gliomas and those without. Since the dataset is labeled, we have thought of using the supervised procedure to address this issue. Along with task classification, we have also carried out some DAT research, feature selection, and feature prioritization. We have used the SHAP method, a subset of the explainable AI (XAI) approach, along with a few filter techniques for feature selection. One could think of the XAI as a wrapper-based framework. The end-to-end pipeline—preprocessing, SMOTE, HHO-based feature selection and tuning, model training, and evaluation—is summarized in Fig. 2.

A. Dataset and Preprocessing

We conducted all experiments using a dataset obtained from the well-known and publicly accessible repository of genome atlas data on TCGA . The dataset was created using information from the TCGA-LGG and TCGA-GBM projects. It includes three clinical factors—Gender, Age at diagnosis, and Race—along with 20 commonly mutated molecular biomarkers, all gathered from 839 patients diagnosed with LGG or GBM. Looking at Table I, we can see that all the predictors are categorical, with the exception of the Age at diagnosis, which is represented as a numerical value.

The molecular characteristics are indicated by values of 0 for not mutated and 1 for mutated, based on the TCGA case number. It's important to highlight that there was no need to use any deletion or imputation techniques, as the dataset utilized in the experiments was complete, with no missing values in any of the attributes (predictor variables).

The dataset contains 24 attributes of 839 patients. Among them, 352 are Glioma patients and 487 are non-Glioma patients. As the data was imbalanced, we balanced the data using the SMOTE function.

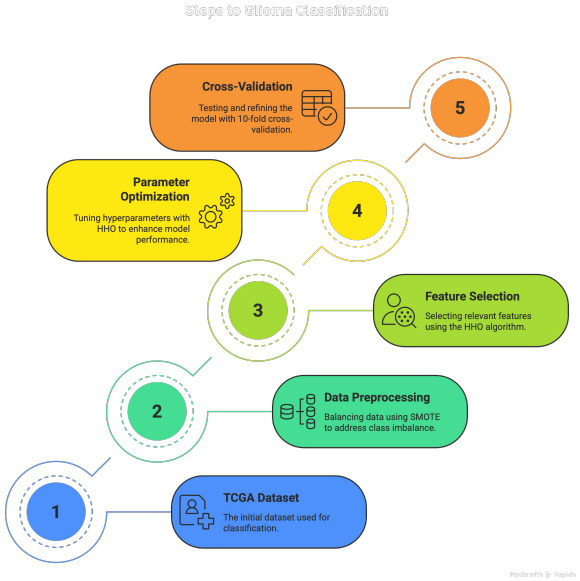


Fig. 2. Methodology for glioma classification using the TCGA dataset. This diagram illustrates the five key steps: dataset selection, data preprocessing with SMOTE, feature selection using HHO, hyperparameter optimization with HHO, and model testing via 10-fold cross-validation.

TABLE I. Predictors in the Dataset Used for Classification

Predictor Name	Category	Possible Values
Gender	Clinical	0, 1
Age	Clinical	[14.42, 89.29]
Race	Clinical	0, 1, 2, 3
IDH1	Molecular	0, 1
TP53	Molecular	0, 1
ATRX	Molecular	0, 1
PTEN	Molecular	0, 1
EGFR	Molecular	0, 1
CIC	Molecular	0, 1
MUC16	Molecular	0, 1
PIK3CA	Molecular	0, 1
NF1	Molecular	0, 1
PIK3R1	Molecular	0, 1
FUBP1	Molecular	0, 1
RB1	Molecular	0, 1
NOTCH1	Molecular	0, 1
BCOR	Molecular	0, 1
CSMD3	Molecular	0, 1
SMARCA4	Molecular	0, 1
GRIN2A	Molecular	0, 1
IDH2	Molecular	0, 1
FAT4	Molecular	0, 1
PDGFRA	Molecular	0, 1

To better understand the interdependence among the features, we calculated the Pearson correlation matrix for all numerical variables in the dataset. The correlation matrix above (Fig. 3) represents, using color and circle size, the effect and direction of linear relationships between features. The blue circles are for negative correlations, the red circles are for positive correlations, and the larger the circle, the stronger the relationship.

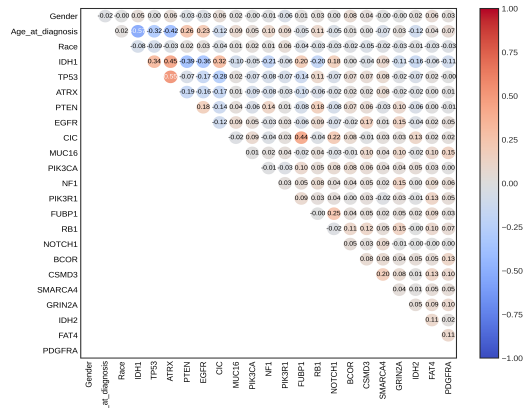


Fig. 3. Upper-triangle correlation matrix of selected clinical and genomic features. Circle size and color intensity indicate the strength and direction of Pearson correlation coefficients (red = positive, blue = negative).

An upper-triangle form such as this one draw attention to potential multicollinearity and redundancies. Specifically, TP53 and IDH1, and Age at diagnosis attributes are moderately correlated ($r2 \geq 0.1$) with multiple other attributes, and therefore may also be interesting predictive features. On the other hand, the lack of strong correlation between most features seems to suggest that this dataset can be used for evaluating and optimizing feature selection separately.

B. Visualization of Data Distribution Using t-SNE

In order to gain insights into the structure of our gene expression dataset and the distribution of our classes within it, we performed dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE). And this is a nonlinear method which embeds the high-dimensional features into a 2D space preserving local similarities so it is very useful for visualizing patterns and class separability. Fig. 4: t-SNE projection of the entire dataset showing class-wise distribution.

Fig. 4. T-SNE projection of the entire dataset showing class-wise distribution.

In Fig. 4, we can see the t-SNE projection of the entire dataset showing Class 0 (purple) and Class 1 (orange) samples in separate, though partially overlapping clusters. This indicates that the input space in high-dimensional has an underlying structure that is probably separable by the proper classifier. This consistency in data distribution across the full dataset validates the efficiency of our preprocessing pipeline (as the distribution of pre-processed data in each set is analogous), and these visualizations make it all the more reassuring.

C. Gradient Boosting Classifier

In this subsection, we discuss the Gradient Boosting classifier used to classify glioblastoma. A classifier is a type of machine learning (ML) algorithm that categorizes data into predefined classes. In this case, the classifier uses patient characteristics, clinical data, to determine whether or not the patient has glioblastoma.

Gradient Boosting (GB) is an advanced ensemble learning algorithm that sequentially constructs weak learners (typically decision trees) to form a strong classifier. It minimizes the prediction error by iteratively adding models that correct the residuals of the previous ensemble. The objective function for GB is:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

Where:

- $L(\theta)$ is the objective function representing the overall loss.
- $l(y_i, \hat{y}_i)$ is the loss function measuring prediction error for each data point.
- $\Omega(f_k)$ is the regularization term to penalize model complexity.

The regularization function is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2}$$

Where:

- T is the number of leaves in the tree.
- w_j is the weight of the j^{th} leaf.
- γ, λ are regularization hyperparameters that control the strength of the regularization.

The model is updated at each iteration using:

$$F_t(x) = F_{t-1}(x) + \alpha h_t(x) \tag{3}$$

Where:

- $F_t(x)$ is the model's prediction after t iterations.
- α is the learning rate controlling the size of updates.
- $h_t(x)$ is the weak learner at iteration t.

D. Performance Metrics

To evaluate the classification performance, several standard metrics were used. These metrics provide insights into the model's ability to correctly identify and classify instances. The key metrics considered include:

- Accuracy: This metric measures the proportion of correctly classified instances out of the total instances. It provides an overall effectiveness of the classification model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- Precision: Precision quantifies the accuracy of the positive predictions made by the model. It is the ratio of true positives to all instances classified as positive, showing how many of the positive predictions were correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- Recall: Recall, also known as sensitivity, measures the model's ability to correctly identify all relevant positive instances. It is the ratio of true positives to the total number of actual positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- F1 Score: The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns. A higher F1 score indicates a better balance between Precision and Recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Additionally, Confusion Matrices and Receiver Operating Characteristic (ROC) Curves were used for visual interpretation of classification performance. The confusion matrix allows for a detailed breakdown of the true vs. predicted classifications, while ROC curves provide insights into the trade-off between the True Positive Rate (Recall) and False Positive Rate at various thresholds.

These metrics offer a comprehensive understanding of the model's classification capabilities and its strengths and weaknesses in distinguishing between classes.

E. Feature Selection Using Harris Hawk Optimization (HHO) Algorithm

Feature selection is the process of finding a noise-free, effective set of features from a given dataset that improves the model's performance on that dataset. The feature selection approach can be broadly divided into three main categories: embedded method, filter-based method, and wrapper-based method. In the wrapper-based method, a subset of the total features is evaluated using a machine learning (ML) algorithm iteratively to find the best feature subset. Metaheuristic algorithms are mostly used in this case to iteratively find an optimum feature subset that provides maximum performance.

To optimize feature selection, we employed the Harris Hawk Optimization (HHO) algorithm—a metaheuristic inspired by the cooperative hunting strategy of Harris hawks. The position of each hawk represents a candidate feature subset. The optimization is driven by exploration and exploitation phases:

- Exploration: Hawks randomly search for solutions based on prey energy.
- Exploitation: Attack strategies like soft besiege, hard besiege, and sudden dives are applied based on the prey's escaping energy.

The energy of the prey is given by:

$$E = 2E_0 \left(1 - \frac{t}{T}\right) \tag{8}$$

Where:

- E_0 is the initial energy of the prey,
- t is the current iteration,
- T is the maximum number of iterations.

Hawks update their position based on the escape energy E using one of the following position update strategies:

$$X(t+1) = X_{\text{prey}}(t) - E \cdot |J \cdot X_{\text{prey}}(t) - X(t)| \tag{9}$$

Where:

- $X(t)$ is the current hawk position,
- $X_{\text{prey}}(t)$ is the position of the prey (best solution),
- J is the random jump strength factor.

For the optimization process, we used the F1 score as the cost function to evaluate the performance of each feature subset. The F1 score, which is the harmonic mean of precision and

recall, was chosen because it provides a balanced evaluation of model performance, especially in imbalanced datasets. The F1 score was computed for each subset, and the feature subset that yielded the highest F1 score was selected as the optimal subset.

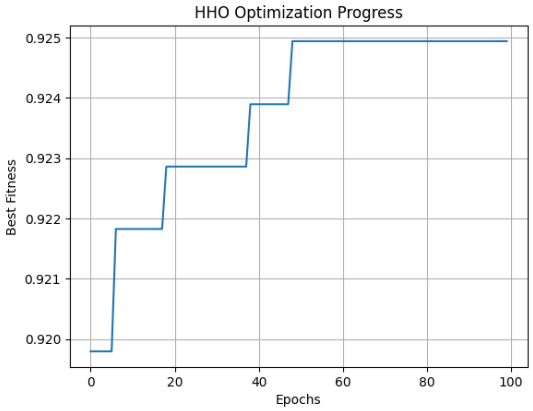


Fig. 5. Harris Hawks Optimization (HHO) convergence curve showing the improvement of best fitness over 100 epochs. The figure illustrates a typical step-wise convergence pattern characteristic of HHO's exploitation-exploration balance.

The Harris Hawks Optimization (HHO) algorithm progressively improves the model's F1 score by selecting optimal feature subsets. As shown in Fig. 5, the F1 score increases in a step-wise manner over the epochs. This reflects HHO's balance between exploration and exploitation, leading to a near-optimal solution.

We initialized 30 hawks over 50 iterations and evaluated subsets using the Gradient Boosting Classifier with 100 estimators. An elitism strategy was used to preserve the top 2 performing hawks during each iteration. The F1 score was maximized by iteratively adjusting the feature set and hyperparameters, ensuring that the final selected features and parameters provided the highest classification performance.

We applied the Harris Hawks Optimization (HHO) algorithm (Algorithm 1) to select an optimal subset of features, which reduced the feature space and improved classification accuracy. The solution sets provided by HHO indicate which features to select. These subsets are then used in the Gradient Boosting Classifier (GBC) for classification, with the F1 score serving as the fitness value for each agent (feature subset). The F1 score effectively balances precision and recall, making it a robust performance metric. In this process, we utilized 30 hawks over 50 iterations. The HHO algorithm updates the hawks' positions based on the best fitness value found at each step. By the end of the process, HHO selects the optimal feature subset, ensuring that the classification model is both efficient and accurate.

Algorithm 1 Feature Selection Using HHO for GBC

Initialize the hawk population $X(i)$ for $i = 1, 2, \ldots, n$
for each hawk in the population **do**
    Decode the feature set from the hawk position
    Calculate the fitness value (F1 score) using the decoded feature set with the Gradient Boosting classifier
end for
Compare the fitness values of each hawk and set X^* as the best hawk
while $t <$ maximum iteration number **do**
    for each hawk in the population **do**
        Update parameters
        if $r < 0.5$ **then**
            if $|A| < 1$ **then**
                Update hawk position towards the global best X^* using the exploration phase
            else
                Select a random position X_{rand}
                Update hawk position towards the random position using the exploration phase
            end if
        else
            Update hawk position towards the global best X^* using the exploitation phase
        end if
    end for
    for each hawk in the population **do**
        Decode the feature set from the hawk position
        Calculate the fitness value (F1 score) using the decoded feature set with the Gradient Boosting classifier
    end for
    Compare the fitness values of each hawk
    if a better solution (feature set) is found **then**
        Update X^*
    end if
end while
Save the best solution set as the final feature set

IV. Results

A. Feature Selection and Hyperparameter Configuration

The feature selection process resulted in the identification of a subset of features that contributed most significantly to the classification performance. The selected features are shown in Table II.

TABLE II. Selected Features After Harris Hawks Optimization

Selected Feature Name			
Gender	Age_at_diagnosis	Race	IDH1
ATRX	PTEN	EGFR	PIK3CA
NF1	SMARCA4	IDH2	PDGFRA

These features were selected based on their ability to differentiate between glioma subtypes. The selection process improved the model's efficiency by reducing the dimensionality of the input data.

The optimal hyperparameters for the Gradient Boosting Classifier (GBC) were determined through the Harris Hawks Optimization (HHO) process. The configurations are shown in Table III.

TABLE III. OPTIMIZED HYPERPARAMETERS FOR THE GBC MODEL

Hyperparameter	Original Configuration	Hyperparameter Tuning	Feature Selection & Tuning
n_estimators	100	144	150
learning_rate	0.1	0.0632	0.0139
max_depth	3	5	5
subsample	1.0	0.8153	0.8465

The optimized values for the hyperparameters show a clear reduction in the learning rate and an increase in the number of estimators, which contribute to the improved performance of the model.

B. Model Performance Comparison

The performance of the model was assessed across three different scenarios: using the original dataset, after hyperparameter tuning, and after both feature selection and hyperparameter tuning. The results, measured using various performance metrics, are summarized in Table IV.

TABLE IV. PERFORMANCE COMPARISON ACROSS DIFFERENT CONFIGURATIONS

Metric	Original Dataset	Hyperparameter Tuning	Feature Selection & Tuning
Accuracy	0.8809 ± 0.0269	0.8830 ± 0.0305	0.8840 ± 0.0266
Precision (Macro)	0.8834 ± 0.0272	0.8846 ± 0.0316	0.8873 ± 0.0264
Recall (Macro)	0.8812 ± 0.0257	0.8826 ± 0.0297	0.8848 ± 0.0253
F1 Score (Macro)	0.8796 ± 0.0265	0.8816 ± 0.0303	0.8829 ± 0.0260
ROC AUC	0.9240 ± 0.0288	0.9304 ± 0.0255	0.9269 ± 0.0315

The results show that feature selection combined with hyperparameter tuning achieved the highest scores across all evaluation metrics. Specifically, the accuracy improved from 0.8809 to 0.8840, while the ROC AUC decreased slightly from 0.9304 to 0.9269 compared to hyperparameter tuning alone, highlighting the effectiveness of feature selection in enhancing model performance. The improvements demonstrate that feature selection, when coupled with hyperparameter tuning, can optimize model performance, confirming its applicability in resource-limited environments.

C. ROC Curve for the Three Models

The Receiver Operating Characteristic (ROC) curve for the best-performing models (Feature Selection & Hyperparameter Tuning, Original Dataset, and HHO Optimization) is shown in Fig. 6. The curves demonstrate the performance of each model, with their respective Area Under the Curve (AUC) values.

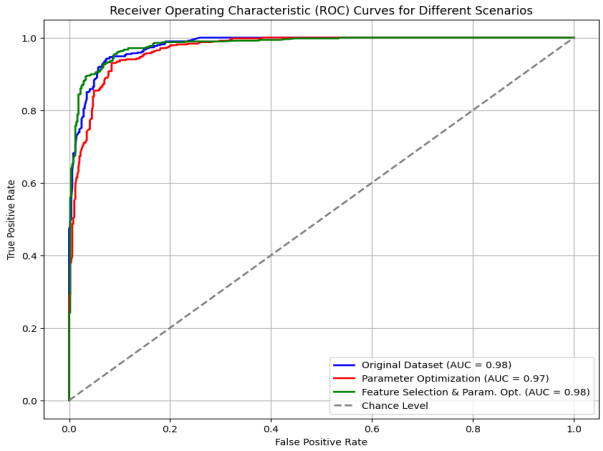


Fig. 6. ROC curve comparison for all three models.

The feature selection + tuning configuration delivers higher accuracy and F1 at the working threshold, while its ROC AUC is slightly lower than the tuning-only configuration. This pattern is consistent with optimizing for F1 rather than global ranking and matches the values reported in Table IV.

D. Confusion Matrix Across Three Models

The confusion matrices for the best fold from three different models—Original Feature Set & Hyperparameter Tuning (OFS & HT), Original Dataset, and Selected Feature Set (SFS) & Hyperparameter Tuning—are presented in Fig. 7a, 7b, and 7c, respectively. These matrices display the key classification metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model. Below are the interpretations drawn from these confusion matrices:

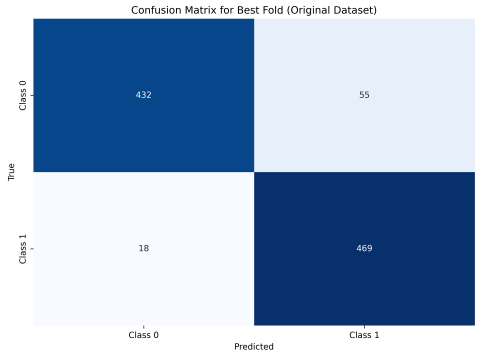
The confusion matrix for the Original Dataset model performs reasonably well, but it reveals a higher number of false positives and false negatives compared to the OFS & HT model. This suggests that, while this model achieves good accuracy, it would benefit from the enhancements offered by feature selection and hyperparameter tuning.

In summary, all models performed well, but the Feature Selection & Hyperparameter Tuning (OFS & HT) model demonstrated the best performance with the highest number of correct classifications. This highlights the importance of feature selection and hyperparameter tuning in achieving optimal classification results.

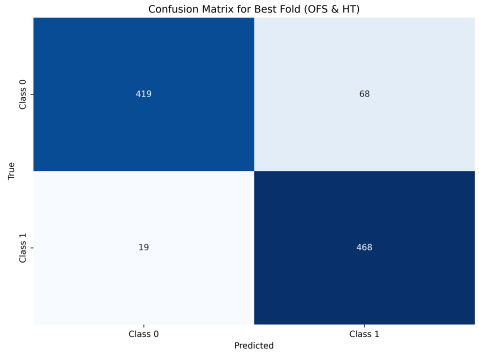
E. K-Fold Cross-Validation Results Across Three Models

Fig. 8 presents the performance metrics across 10-fold cross-validation for each of the three models. Boxplots show the spread and central tendency of each metric, including accuracy, precision, recall, F1 score, and ROC AUC, providing a comparison of model stability and performance.

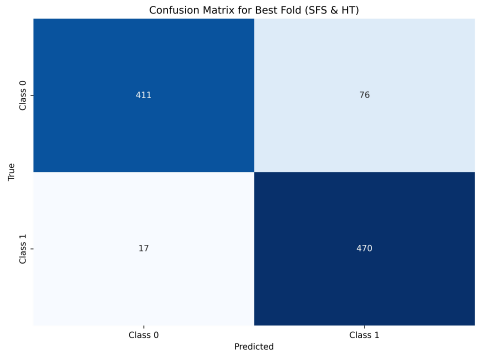
These boxplots highlight that the SFS & HT model is the best-performing model, exhibiting consistent and superior performance across all evaluation metrics.



(a) Confusion matrix for original dataset.



(b) Confusion matrix for hyperparameter tuning only.

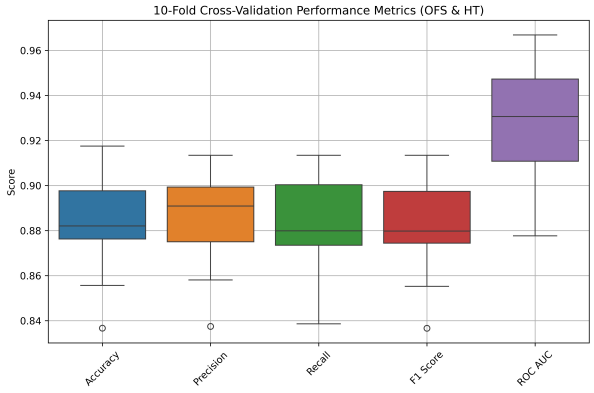


(c) Confusion matrix for feature selection + tuning.

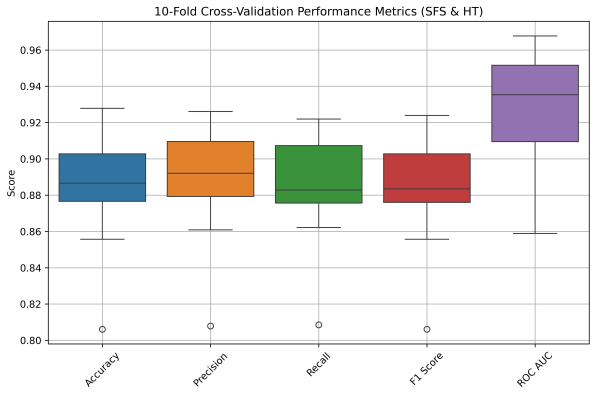
Fig. 7. Confusion matrices for the best fold across three models: Original dataset, hyperparameter tuning only, and feature selection + tuning.

F. SHAP-Based Validation of Selected Features

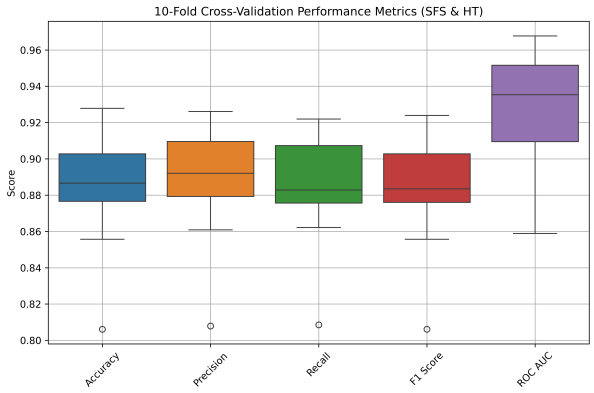
We used SHAP (SHapley Additive exPlanations) to evaluate the output of the trained model on the training data, thereby validating the resilience of the feature subset chosen using the Harris Hawks Optimization (HHO) algorithm. SHAP, based on cooperative game theory, offers consistent, locally accurate feature attributions, making it a useful instrument for post-hoc model interpretability. By providing a clear understanding of how each feature contributes to model predictions, SHAP helps validate the feature selection process and ensures the reliability of the model's decisions.



(a) Original dataset.



(b) Hyperparameter tuning.



(c) Feature selection + tuning.

Fig. 8. 10-Fold cross-validation performance metrics for the three experimental models: Original dataset, hyperparameter tuning only, and feature selection + hyperparameter tuning. Metrics include accuracy, precision, recall, F1 score, and ROC AUC.

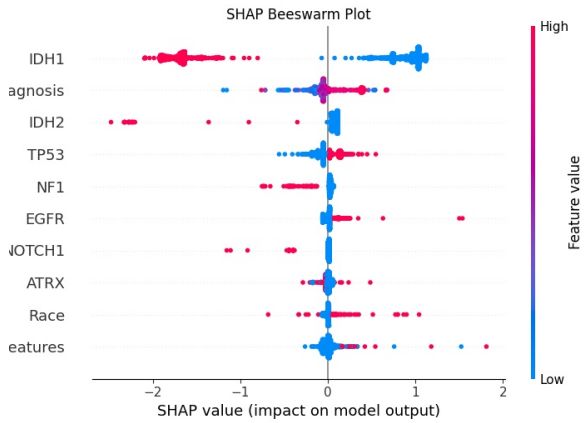


Fig. 9. SHAP summary plot illustrating the impact of individual features on the model's output for the training set. Red and blue represent high and low feature values, respectively.

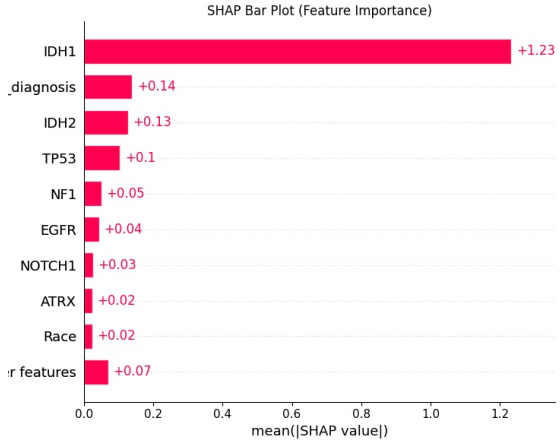


Fig. 10. Mean absolute SHAP values indicating the overall importance of features across all training samples. Top-ranked features include IDH1, Age_at_diagnosis, PTEN, and IDH2.

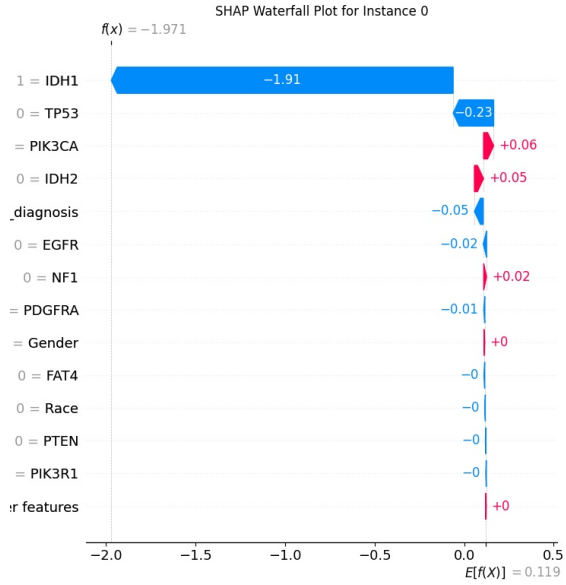


Fig. 11. SHAP waterfall plot illustrating individual feature contributions for a specific prediction. The plot shows the cumulative impact of each feature.

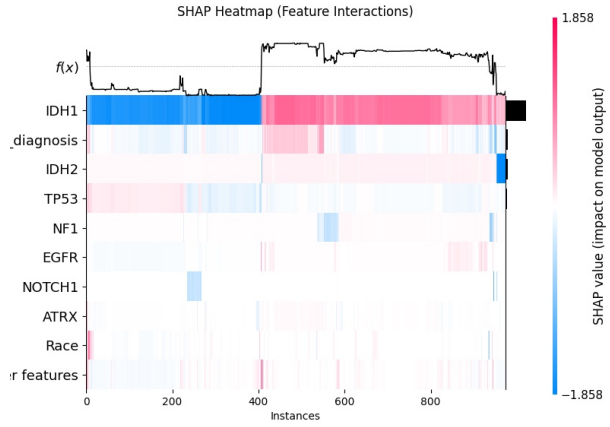


Fig. 12. SHAP heatmap showing feature importance and interactions across all predictions.

Fig. 9 illustrates the SHAP summary plot, which shows the impact of individual features on the model's output for the training set. Red and blue represent high and low feature values, respectively. The most powerful features influencing model predictions are revealed by the SHAP summary plot and the bar plot of mean absolute SHAP values (Fig. 10). Not least among these are IDH1, Age_at_diagnosis, PTEN, IDH2, EGFR, TP53, and NF1. These traits have great predictive value since they show as high-impact contributors over several samples.

The SHAP waterfall plot in Fig. 11 illustrates individual feature contributions for a specific prediction. The plot displays the cumulative impact of each feature, helping us understand how the model arrived at its output for a particular instance.

Finally, Fig. 12 shows the SHAP heatmap, which illustrates feature importance and interactions across all predictions. It provides a visual representation of how features interact and their relative importance in the overall model.

Especially, several of the top SHAP-ranked features are included in the subset of features chosen by the HHO algorithm — `['Age_at_diagnosis', 'IDH1', 'EGFR', 'NF1', 'PIK3R1', 'FUBP1', 'NOTCH1', 'BCOR', 'IDH2', 'FAT4']` —. Particularly for IDH1, Age_at_diagnosis, EGFR, NF1, IDH2, NOTCH1, and BCOR, the overlap between the SHAP analysis and HHO selection offers independent proof that the optimization procedure effectively found highly impactful factors for the model.

This agreement suggests that the HHO-based feature selection technique captured features of actual relevance for classification, thereby supporting the validity of this approach. The general interpretability and resilience of the proposed framework are supported by the agreement between model explanation and optimization results.

We tuned the model to work best at one practical decision point (the F1 balance), not across every possible threshold. That is why accuracy and F1 increased while ROC AUC, which reflects overall ranking across all thresholds, dipped a little: some score ordering changed, but decisions at the chosen cutoff improved. The SHAP results show that a small set of biologically sensible features—IDH1, EGFR, TP53, NF1, and Age—push predictions in expected directions; "high-risk" patterns raise the score, while "lower-risk" patterns reduce it. In practice, a high score with a few strong drivers can prompt faster confirmatory testing or earlier team review, whereas a low score with mixed or weak drivers can follow routine care. This ties the numbers back to clinical meaning and explains why the confusion matrices show fewer mistakes at the working threshold.

V. Discussion

This study shows that combining Harris Hawks Optimization (HHO) with a Gradient Boosting Classifier (GBC) can produce a compact and accurate model for glioma classification while keeping the reasoning behind predictions understandable. In our view, the most encouraging aspect is that the features highlighted by the model—such as IDH1, EGFR, TP53, NF1, and Age—are consistent with the biological story clinicians expect to see in glioma. This agreement suggests the model is learning meaningful patterns rather than overfitting to noise, and it makes the outputs easier to discuss in multidisciplinary settings.

We also observe a clear trade-off between threshold-based metrics and ranking-based metrics. After feature selection and tuning, accuracy and F1 improved, while ROC AUC decreased slightly. This is a reasonable outcome because the optimization process concentrated on improving the balance of precision and recall at a working threshold, not on maximizing performance across all possible thresholds. In practical terms, the model became better at identifying true cases at an operating point that matters for care, even if overall ranking changed a little. Because clinical use often depends on a specific decision threshold, we consider this an acceptable and transparent trade-off.

Another practical strength is parsimony. The HHO search consistently pushed the model toward a smaller set of informative variables, which reduces redundancy, speeds up training and inference, and lowers the cost of deployment. The resulting explanations from SHAP are easier to interpret at the patient level because fewer features dominate the prediction. In a clinical workflow, this can help frame conversations such as why the model flagged a case and which factors were most influential at that moment.

Finally, we see a realistic path to use: align the decision threshold with local practice, check probability calibration, review explanations alongside routine clinical information, and monitor performance over time. The goal is not to replace clinical judgment but to support it with a tool that is fast, consistent, and explainable.

VI. Limitations

This work uses a single public dataset for development and internal testing. Although it includes many patients, it may not capture the full diversity of real-world populations, imaging or sequencing methods, and clinical practice. As a result, the model's performance could change when applied to other hospitals or regions. External validation on independent cohorts from different institutions is needed to confirm generalizability.

The labeling scheme and endpoint definition may not fully match current clinical reporting standards. While the model distinguishes classes effectively within the dataset, future versions should align labels more closely with contemporary diagnostic categories to better reflect how decisions are made in practice.

Class imbalance was addressed with synthetic oversampling during training, which can influence the distribution and calibration of predicted probabilities. Although this helped improve recall and F1 at the chosen threshold, probability estimates may require calibration before the model is used to trigger clinical actions. Threshold choice should also reflect local risk tolerance and downstream resource constraints.

Metaheuristic optimization introduces randomness through initialization and fold splits. Although we used cross-validation and stable settings, minor variation between runs is expected. A more exhaustive stability assessment across multiple seeds and repeated folds would provide stronger evidence that the selected feature set and hyperparameters are robust.

Lastly, while SHAP improves transparency, it is a post-hoc explanation technique and does not prove causality. Some interactions between features may be complex and context-dependent, and explanations should be interpreted as supportive evidence rather than definitive biological mechanisms. Future work will focus on external testing, probability calibration, threshold setting with clinical input, and prospective evaluation in a workflow that tracks utility and safety over time.

VII. Conclusions

This study demonstrates that the Gradient Boosting Classifier (GBC), optimized using Harris Hawks Optimization (HHO) for feature selection and hyperparameter tuning, significantly enhances glioma classification performance. Despite reducing the feature set by 10 features, the Feature Selection & Hyperparameter Tuning (OFS & HT) model still outperforms other models in terms of accuracy, precision, recall, and F1 score. By selecting a more compact set of features, the model not only delivers better results but also becomes more efficient and cost-effective; fewer inputs translate into lower computational cost and faster processing, which is practical for real-world use. SHAP analysis further validates the importance of the selected features—such as IDH1, EGFR, and NF1—and, importantly, makes each prediction explainable, so that clinicians can see why a case was flagged and discuss the reasoning in tumor-board or radiology review.

Beyond technical metrics, the intended clinical value is early and reliable support: high-confidence outputs with clear drivers can prompt faster biomarker confirmation and earlier escalation of complex cases, while low-risk outputs with diffuse drivers can proceed through routine pathways. To ensure that these benefits hold outside of development data, we will perform external validation on independent cohorts

from other institutions and time periods, calibrate predicted probabilities, and select operating thresholds with clinical input. We also plan a prospective pilot in routine workflow to measure impact on turnaround time, downstream testing, and patient management, while monitoring stability across seeds, splits, and patient subgroups. In this way, the system moves from strong performance on paper toward safe, useful, and sustainable deployment in practice, while maintaining efficiency and interpretability.

REFERENCES

[1] World Health Organization, "International Agency for Research on Cancer," https://gco.iarc.fr, n.d..

[2] Cleveland Clinic, "Glioma," https://my.clevelandclinic.org/health/diseases/21969-glioma, n.d..

[3] Miller, K.D., Siegel, R.L., Lin, C.C., Mariotto, A.B., Kramer, J.L., Rowland, J.H., Stein, K.D., Alvaro, M., Jemal, A., "Cancer statistics for adolescents and young adults, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 6, pp. 443–459, 2020.

[4] Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., the WHO Classification of Tumours of the Central Nervous System, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, no. , pp. 97–109, 2007.

[5] Gaillard, F., "WHO Classification of CNS Tumors," https://radiopaedia.org/articles/who-classification-of-cns-tumours-1?lang=us, 2022.

[6] Hanif, F., Dastgir, G., Abbas, Q., Hussain, I., Ghaffar, S., "Glioblastoma multiforme: A review," *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 1, pp. 3–9, 2017.

[7] Mirchia, K., Richardson, T., "Beyond IDH-mutation in gliomas," *Cancers*, vol. 12, no. , pp. 1817, 2020.

[8] Vigneswaran, K., Esquenazi, Y., Melkonian, S.C., Chow, A., Chapman, M., Green, S., Min, B., Nasrallah, A., Lanchon, S., Pradhan, R., "Advances in glioma genetics," *Annals of Translational Medicine*, vol. 3, no. , pp. 95, 2015.

[9] DeWitt, J., Choi, M., Schwartz, A., Epstein, M., Jolly, M., Mishra, D., "Cost-effectiveness of IDH testing in gliomas," *Neuro-Oncology*, vol. 19, no. , pp. 1640–1650, 2017.

[10] Krauze, A., Manley, M., Jang, C., Olson, R., Saeed, R., Shukla, N., Das, S., "AI-Driven Image Analysis in CNS Tumors," *Journal of Biotechnology and Biomedicine*, vol. 5, no. , pp. 1–19, 2022.

[11] Diaz Rosario, M., Tohme, R., Islam, N., "Sex Differences in Glioma Data," *Biomolecules*, vol. 12, no. , pp. 1203, 2022.

[12] Liu, X., Zhang, L., Zhang, S., Wu, X., "Deep Learning in Glioma Prognosis: A Survey," *Frontiers in Oncology*, vol. 10, no. , pp. 798, 2020.

[13] Ko, C., Brody, J. P., "A genetic risk score for glioblastoma multiforme based on copy number variations," *Cancer Treatment Research Communications*, vol. 27, no. , pp. 100352, 2021.

[14] Sánchez-Marqués, R., García, V., Sánchez, J. S., "A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data," *Scientific Reports*, vol. 14, no. , pp. 17195, 2024.

[15] Kha, Y.-H., others, "Development and validation of an efficient MRI radiomics signature for improving the predictive performance of 1p/19q codeletion in lower-grade gliomas," *Cancers (Basel)*, vol. 13, no. 21, pp. 5398, 2021.

[16] Xia, X., others, "Interpretable Machine Learning Models for differentiating glioblastoma from solitary brain metastasis using radiomics," *Academic Radiology*, vol. , no. , pp. , to be published, 2025.

[17] Tang, Y., others, "Identification of five important genes to predict glioblastoma subtypes," *Neuro-Oncology Advances*, vol. 3, no. 1, pp. , 2021.

[18] Kumar, D., others, "An improved machine-learning approach for COVID-19 prediction using Harris Hawks Optimization and feature analysis using SHAP," *Diagnostics*, vol. 12, no. 5, pp. 1023, 2022.

[19] Nalasari, L. T., Anam, S., Shofianah, N., "Liver cirrhosis classification using extreme gradient boosting classifier and Harris Hawks optimization as hyperparameter tuning," *Journal of Electronics, Electromedical Engineering and Medical Informatics*, vol. 7, no. 2, pp. , 2025.

[20] Kurdi, S. Z., others, "Brain tumor classification using meta-heuristic optimized convolutional neural networks," *Journal of Personalized Medicine*, vol. 13, no. 2, pp. 181, 2023.

[21] Pirgazi, J., others, "An efficient hybrid filter-wrapper method based on improved Harris Hawks optimization for feature selection," *BioImpacts*, vol. , no. , pp. , 2024.

[22] Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., Soffietti, R., von Deimling, A., Ellison, D.W., "The 2021 WHO classification of tumours of the central nervous system," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021.