Integrating explainable artificial intelligence and light gradient boosting machine for glioma grading

Article · December 2024

DOI: 10.1016/j.infoh.2024.12.001

Teuku Rizky Noviandy
Universitas Abulyatama
75 PUBLICATIONS **1,018** CITATIONS

SEE PROFILE

Ghalieb Mutig Idroes
Graha Primera Saintifika
40 PUBLICATIONS **671** CITATIONS

SEE PROFILE

Irsan Hardi
Graha Primera Saintifika
36 PUBLICATIONS **649** CITATIONS

SEE PROFILE

Contents lists available at ScienceDirect

Informatics and Health

journal homepage: www.keaipublishing.com/en/journals/informatics-and-health/

Integrating explainable artificial intelligence and light gradient boosting machine for glioma grading

Teuku Rizky Noviandy[a,b,*], Ghalieb Mutig Idroes[b], Irsan Hardi[b]

[a] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia
[b] Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar 23771, Indonesia

ARTICLE INFO

Keywords:
LightGBM
SHAP
XAI
Model Interpretability
Machine Learning
Classification

ABSTRACT

Background: Glioma grading plays a pivotal role in neuro-oncology, directly influencing treatment strategies and patient prognoses. Despite its importance, traditional histopathological analysis has drawbacks, spurring interest in applying machine learning (ML) techniques to improve accuracy and reliability in glioma grading.
Methods: This study employs the Light Gradient Boosting Machine (LightGBM), an advanced ML algorithm, in combination with Explainable Artificial Intelligence (XAI) methodology to grade gliomas more effectively. Utilizing a dataset from The Cancer Genome Atlas, which comprises molecular and clinical characteristics of 839 glioma patients, the LightGBM model is meticulously trained, and its parameters finely tuned. Its performance is benchmarked against various other ML models through a comprehensive evaluation involving metrics such as accuracy, precision, recall, and F1-score.
Results: The optimized LightGBM model demonstrated exceptional performance, achieving an overall accuracy of 89.88 %, which surpassed the other compared ML models. The application of XAI techniques, particularly the use of Shapley Additive Explanation (SHAP) values, revealed the IDH1 gene mutation as a significant predictive factor in glioma grading, alongside providing valuable insights into the model's decision-making process.
Conclusions: The integration of LightGBM with XAI techniques presents a potent tool for glioma grading, showcasing high accuracy and offering interpretability, which is crucial for gaining clinical trust and facilitating broader adoption. Despite the promising results, the study acknowledges the need to address dataset limitations and the potential benefits of incorporating a more comprehensive range of features in future research to refine and further enhance the model's applicability and performance in clinical settings.

1. Introduction

Glioma is a type of brain tumor that arises from glial cells and represents a significant challenge in the field of neuro-oncology.[1] These tumors exhibit a wide range of aggressiveness and heterogeneity, making accurate grading essential for treatment planning and prognosis. Gliomas can vary from low-grade tumors with a more favorable prognosis to high-grade tumors associated with increased morbidity and mortality.[2,3]

Traditional glioma grading relies on histopathological analysis, which involves examining tissue samples under a microscope.[4] While this approach has been the cornerstone for decades, it has limitations, including subjectivity, intra- and inter-observer variability, and time-consuming processes.[5] Moreover, the static nature of histopathological assessments may not capture the dynamic changes within tumors, hindering real-time decision-making in clinical settings.[6]

In recent years, the integration of machine learning (ML) techniques has shown promise in enhancing the accuracy and efficiency in various tasks.[7–10] ML algorithms can analyze complex patterns within medical data, providing a quantitative and objective means of tumor characterization. This shift towards data-driven approaches opens new avenues for improving diagnostic precision and personalized treatment strategies.[11] Among the diverse ML algorithms, the Light Gradient Boosting Machine (LightGBM) has gained prominence for its exceptional performance across various tasks, such as drug discovery, medical diagnostics, and disease prediction, demonstrating its versatility and effectiveness.[12–15]

Explainable Artificial Intelligence (XAI) is an approach that aims to make AI-driven decisions clear and understandable.[16] As machine learning models become more complex and their outputs increasingly

[*] Corresponding author at: Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia.
E-mail address: trizkynoviandy@gmail.com (T.R. Noviandy).

https://doi.org/10.1016/j.infoh.2024.12.001
Received 5 December 2023; Received in revised form 27 October 2024; Accepted 12 December 2024
Available online 24 December 2024

impact patient care, XAI has become essential in healthcare. XAI methods help healthcare providers see how specific model predictions are made, breaking down which features contribute to particular decisions. This transparency is crucial in clinical settings, where decisions carry high stakes and need to be understandable and trustworthy.[17–19] By clarifying model reasoning, XAI supports clinicians in making informed, evidence-based decisions, which helps integrate AI tools more effectively into patient care.

In parallel with these developments, the demand for interpretability in healthcare ML applications has risen. Shapley Additive Explanations (SHAP), a widely used XAI method grounded in cooperative game theory, is particularly valuable in this context.[20] SHAP assigns importance values to features in a model, providing a comprehensive and mathematically rigorous explanation of feature contributions to predictions. This method's appeal lies in its ability to offer global interpretability, fostering clinician trust and supporting the integration of ML into neuro-oncological practice by enhancing transparency in clinical decision-making.[21,22]

Despite the advancements in ML applications in healthcare, there exist research gaps in the specific context of glioma grading. The current limitations of traditional histopathological analysis highlight the need for more robust approaches. The need for interpretability poses a significant challenge to their acceptance in clinical settings. Bridging these gaps is essential for realizing the full potential of ML in improving decision-making processes in neuro-oncology.Top of Form

In this study, our primary objective is to enhance the interpretability and transparency of glioma grading tasks by integrating the LightGBM with XAI represented by the SHAP method. The selection of LightGBM is motivated by its established effectiveness in diverse tasks, efficient handling of large datasets, and notable speed, rendering it an optimal choice for automating glioma grading tasks.[23–25] Its demonstrated versatility and robust performance, particularly when compared to deep neural networks for tabular data, align well with the characteristics of the medical dataset employed in this study.[26–28] Regarding the SHAP method, it was chosen due to its unique capability to offer global interpretability.[29] This means that the SHAP method enables us to understand the importance of features across the entire dataset.

The study contributed by integrating a powerful ML algorithm with an interpretable and transparent XAI method that aimed to provide clinicians with not only accurate predictions but also a comprehensible understanding of the factors influencing those predictions. The combination of LightGBM and SHAP is expected to enhance the diagnostic precision of glioma grading, offering a quantitative and objective framework for personalized treatment strategies. The study's contributions extended beyond the technical aspects, as it addressed the crucial need for trust and transparency in the integration of ML technologies into clinical practice.

2. Related works

The incorporation of ML techniques in glioma grading processes represents a significant advancement. It denotes a notable change in how gliomas are assessed and classified, potentially leading to more accurate diagnoses and better treatment decisions. Shboul et al. achieved significant results in the non-invasive prediction of molecular mutations in low-grade gliomas using ML models developed for this purpose, which showed notable test performance in terms of area under the curve (AUC). The AUC for MGMT methylation was 0.83 ± 0.04, for IDH mutation was 0.84 ± 0.03, for 1p/19q co-deletion was 0.80 ± 0.04, for ATRX mutation was 0.70 ± 0.09, and for TERT mutations was 0.82 ± 0.04.[30] These findings underscore the potential of ML in predicting clinically relevant molecular markers, which can guide treatment decisions and improve patient outcomes. However, the study's limitations, such as the need for larger and more diverse datasets, warrant further investigation to ensure the generalizability of the models.

How

In a separate study, Wu et al. focused on identifying the optimal radiomics-based ML method for predicting the isocitrate dehydrogenase (IDH) genotype in diffuse gliomas. They evaluated eight classical ML methods, assessing their stability and performance. The results of their study were particularly noteworthy for the Random Forest algorithm. This method demonstrated high predictive performance, with an accuracy of 0.885 ± 0.041 and an area under the curve (AUC) of 0.931 ± 0.036.[31] The superior performance of Random Forest compared to other ML methods highlights the importance of selecting the most suitable algorithm for specific prediction tasks. However, the study could benefit from further exploration of the clinical implications of IDH genotype prediction and its role in guiding treatment strategies.

In another study, Tasci et al. introduced a novel hierarchical voting-based methodology to enhance the performance of feature selection stages and ML models in glioma grading, integrating clinical and molecular features. This approach involved using publicly available datasets from The Cancer Genome Atlas (TCGA) and the Chinese Glioma Genome Atlas (CGGA). The computational results of this innovative approach were impressive: the proposed method achieved accuracy rates of 87.606 % on the TCGA dataset and 79.668 % on the CGGA dataset. These results notably outperformed those achieved by the LASSO feature selection method alone.[32] The proposed methodology's ability to effectively integrate diverse features and its robustness across different datasets underscore its potential for improving glioma grading. However, a more in-depth analysis of the specific advantages and limitations of the hierarchical voting-based approach could provide valuable insights.

Despite the promising results of these studies, all share a common limitation that requires improvement: interpretability. The lack of interpretability in the employed ML models poses significant challenges in clinical settings, hindering clinicians' trust and understanding of the predictions, which is important for informed decision-making. The complex nature of ML models often makes it difficult for healthcare professionals to fully embrace and utilize their outputs, particularly when the rationale behind each prediction can significantly impact treatment planning and patient outcomes.

3. Methods

The workflow of this study is presented in Fig. 1. We utilized a dataset to train the LightGBM model, which was trained using cross-validation and optimized through hyperparameter tuning. The model was then employed to predict outcomes on the testing set, and the results were made interpretable using the SHAP method.

3.1. Dataset

The dataset utilized in this study was obtained from prior research conducted by Tasci et al.[32] This dataset is constructed based on the comprehensive genomic and clinical information available in TCGA datasets, specifically TCGA-Lower Grade Glioma (TCGA-LGG) and TCGA-Glioblastoma Multiforme (TCGA-GBM) brain glioma projects. The dataset comprises records from 839 patients diagnosed with brain glioma. Each patient record is characterized by a combination of clinical and molecular features, providing a holistic view of the tumor and the patient's demographics. Three clinical features are gender, age, and race, and 20 molecular features. All features have been preprocessed to possess numerical values.

The molecular features consist of IDH1, TP53, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, GRIN2A, IDH2, FAT4, and PDGFRA. Each molecular feature is encoded as 0 for not mutated or 1 for mutated. Similarly, gender is encoded with 0 representing male and 1 representing female. For race, the encoding is as follows: 0 for White, 1 for Black or African American, 2 for Asian, and 3 for American Indian or Alaska Native.

Within the dataset, glioma grading is categorized into two types:

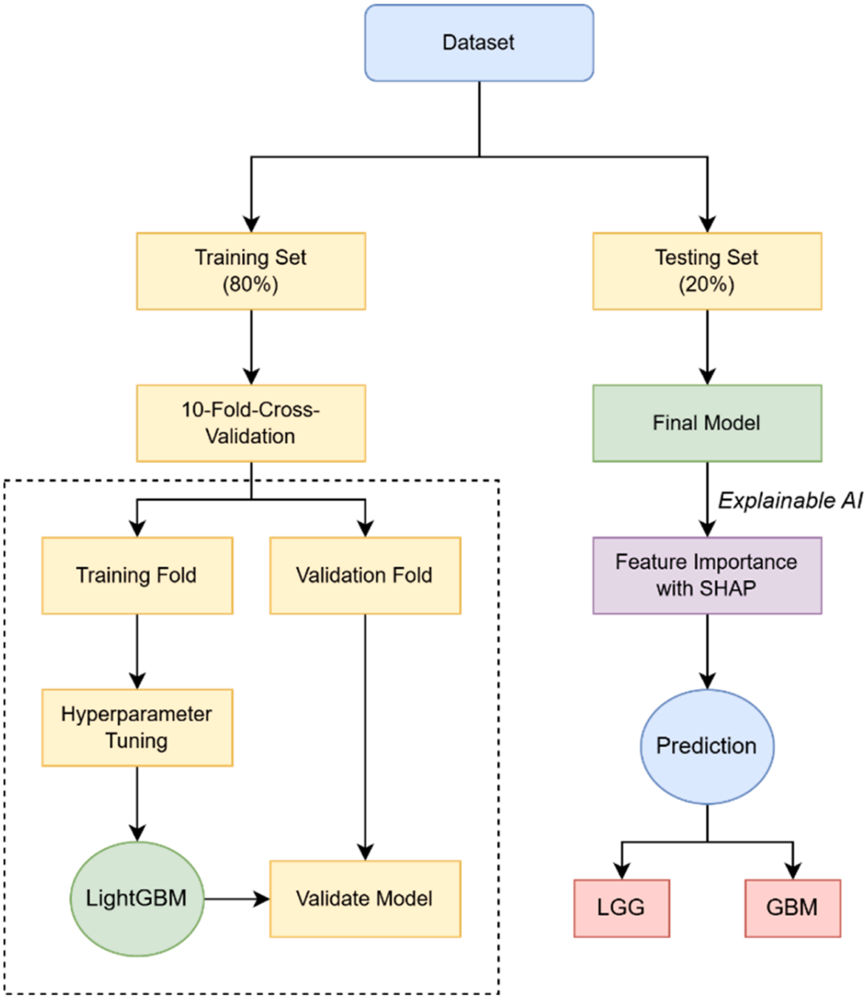

Fig. 1. Workflow of this study.

lower-grade glioma (LGG) and glioblastoma multiforme (GBM). The dataset consists of 487 instances of LGG and 352 instances of GBM. This slight imbalance in class representation reflects the heterogeneity of gliomas, encompassing both less aggressive and more aggressive forms, and ensures adequate representation for training and evaluating the ML model

3.2. Model training

The LightGBM model was chosen for its efficiency and effectiveness in handling tabular data for classification tasks. It utilized a novel "leaf-wise" tree growth approach that prioritized leaf nodes with the maximum loss reduction.[33,34] This strategy led to faster training times and efficient dataset handling compared to traditional gradient boosting methods. LightGBM also effectively addressed class imbalance through techniques such as gradient-based sampling and tree-based learning.

Gradient-based sampling focuses on selecting instances for training based on their gradient, which reflects how much the prediction error would change if the model's predictions were adjusted.[35] By emphasizing instances with larger gradients, the model ensures that it learns more effectively from the most challenging cases, which is crucial when dealing with imbalanced datasets like those involving LGG and GBM cases.

Tree-based learning in LightGBM constructs decision trees that are capable of capturing complex data patterns. This method allows the model to handle imbalances by growing trees that focus on reducing misclassification errors, particularly for the minority class.[36] As a result, LightGBM can more accurately predict glioma grades despite the challenges posed by an imbalanced dataset. These combined capabilities align well with the study's aim of achieving precise and reliable glioma classification.

Before initiating model training, the dataset was partitioned into training and testing subsets using an 80–20 split, where 80 % of the data is allocated for training the model and the remaining 20 % for evaluating its performance.[37] During training, the model is provided with the training set comprising both clinical and molecular features as input, along with the corresponding glioma grades as labels. The LightGBM algorithm iteratively learns the underlying patterns in the data by minimizing a predefined loss function through gradient boosting.

3.3. Model optimization

In order to enhance the performance of the LightGBM model for glioma grading, hyperparameter tuning was conducted.[38] Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a ML model.[39] Hyperparameters are configuration settings that are not learned from the data but are set before the training process. Tuning involves searching through different combinations of hyperparameter values to enhance the model's performance.[40,41]

The tuning method used in this study was a random search approach. Random search is a hyperparameter tuning method where a random set of combinations is selected and evaluated.[42] This approach is more computationally efficient when dealing with a large hyperparameter space. By randomly sampling different combinations, it increases the likelihood of finding good hyperparameter values without testing every possible combination.

The random search conducted with 10-fold cross-validation involves dividing the dataset into 10 folds, training the model on 9 folds, and validating the remaining fold.[43] This process is repeated ten times, ensuring that each fold serves as the validation set exactly once. The average accuracy across all folds is then used to evaluate the model's overall performance, and the best model is selected with the highest accuracy after 100 iterations. The hyperparameter space explored during the random search is presented in Table 1. The hyperparameters considered include the number of trees, the maximum depth of each decision tree, the learning rate, the subsample fraction, and the colsample_bytree.

3.4. Performance Evaluation

A comprehensive model evaluation was conducted to assess the performance of the LightGBM model for glioma grading. The evaluation involved a comparison with five other ML models: Random Forest, Naïve Bayesian, Decision Tree, Support Vector Machine, and k-Nearest Neighbors.

The evaluation metrics utilized for comparison encompass accuracy, precision, recall, and F1-score. Accuracy functions as a measure of the models' correctness.[44,45] Precision reflects the proportion of true positive predictions among all positive predictions made by the model, indicating its ability to avoid false positives. Recall, on the other hand, measures the proportion of true positive predictions captured by the model among all actual positive instances, highlighting its ability to detect all positive instances effectively. The F1-score, as the harmonic mean of precision and recall, furnishes a balanced performance measure. In this study, LGG represents true positive instances, and GBM represents true negative instances. The equations for accuracy, precision, recall, and F1-score are presented in Eqs. 1–4: [46]

$$Accuracy = \frac{TP + FN}{FP + FN + TP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where TP (True Positive) refers to the number of LGG cases that are correctly predicted as LGG, TN (True Negative) represents the number of GBM cases that are accurately predicted as GBM. FP (False Positive) denotes the number of GBM cases incorrectly classified as LGG, and FN (False Negative) is the number of LGG cases that are mistakenly predicted as GBM.

Furthermore, Receiver Operating Characteristic (ROC) curves were utilized to assess the discriminative capabilities of each ML model for glioma grading. These curves provide visual representations of the trade-off between true positive rate and false positive rate across different classification thresholds.[47] Additionally, the AUC score was calculated for each model, serving as a quantitative measure of its discriminatory power. The AUC values obtained from the ROC curve analysis offering insights into the models' abilities to differentiate between LGG and GBM cases. A higher AUC indicates superior performance in distinguishing between these tumor types, which holds significant importance for precise diagnosis and subsequent treatment planning in clinical settings.

Table 1
Hyperparameter space used in this study.

Hyperparameter	Description	Values
n_estimators	The number of trees added sequentially during boosting	A random integer between 50 and 500
max_depth	Maximum depth of each decision tree	A random integer between 3 and 15
learning_rate	The scaling factor for the contribution of each tree	100 values evenly spaced between 0.01 and 0.3
subsample	Fraction of data used for training each tree	100 values evenly spaced between 0.2 and 1.0
colsample_bytree	Fraction of features randomly sampled for each tree	100 values evenly spaced between 0.2 and 1.0

3.5. Model interpretation

We utilized XAI principles through the integration of SHAP for the interpretation of the LightGBM model used in glioma grading. SHAP is a game theory-based approach that explains the output of ML models. It assigns each feature an importance value for a particular prediction, providing a detailed view of how each feature contributes to the model's decision-making process.[48,49] These importance values, quantify the contribution of each feature to the predicted outcome for each individual data point. A positive SHAP value indicates that the feature pushes the model's prediction toward a higher probability of a certain outcome, while a negative SHAP value pushes it toward a lower probability.

SHAP values were calculated for each feature in our dataset. This was achieved by aggregating the SHAP values across all data points to determine the most influential features in predicting glioma grades. By aggregating SHAP values across all data points, we identified the most influential features in predicting glioma grades. Features with larger absolute SHAP values exert a stronger influence on the model's predictions, making them critical for understanding the decision-making process. This detailed analysis allows for a comprehensive understanding of how each feature contributes to the model's output, offering transparency and interpretability in the context of glioma grading.

3.6. Experimental setup

This study was conducted on a computer equipped with hardware including an Intel i5 12400 F processor running at 2.50 GHz and 16 GB of RAM. The code was written in Python v3.10.9. The LightGBM model was implemented using version 4.1.0 of the LightGBM library, while version 1.2.0 of the scikit-learn library was utilized for additional ML functionalities. SHAP analysis was performed using version 0.41.0 of the SHAP library. To ensure the reproducibility of results, all random states used throughout the experiments were set to 42. This standardized approach enables the replication of results and facilitates comparisons across different experiments or studies.

4. Results

4.1. Performance analysis

In this study, we successfully trained LightGBM for glioma grading, achieving notable results in accurately classifying glioma grades. Through a rigorous hyperparameter tuning process, specific parameter values were selected to enhance the model's performance. The hyperparameters for the trained LightGBM model are as follows: n_estimators set to 450, max_depth set to 5, learning_rate set to 0.028, subsample set to 0.46, and colsample_bytree set to 0.26. These parameters were meticulously determined through a random search with 10-fold cross-validation, aiming to improve the model's accuracy and effectiveness in glioma grading. Notably, the combination of these hyperparameters indicates a balanced model architecture: a relatively high number of estimators suggests robustness in capturing complex patterns in the data, while moderate values for max_depth, learning_rate, subsample, and colsample_bytree ensure avoidance of overfitting and efficient utilization of features during training.

The comparative analysis of ML models' performance for glioma grading is presented in Table 2. The optimized LightGBM model demonstrated an impressive overall accuracy of 89.88 %, outperforming other ML models considered in the comparative analysis. This figure highlights its exceptional capacity to accurately classify glioma grades compared to its counterparts. The high accuracy attributed to the LightGBM model implies its effectiveness in minimizing both false positives and false negatives, thereby establishing a robust framework for precise glioma grading.

While LightGBM achieves a commendable precision of 96.67 %, it slightly trails behind the Naïve Bayesian model, which boasts a precision of 98.75 %. The Naïve Bayesian model's marginally higher precision suggests its potential for accurately identifying LGG without mistakenly labeling GBM cases. In terms of recall, both the Random Forest and Decision Tree models surpass LightGBM, achieving a recall rate of 88.12 %. The superior recall rates of these models indicate their potential for correctly identifying LGG cases, although their overall balance with metrics like precision and accuracy requires consideration.

Although LightGBM does not individually lead in precision or recall, its F1-score suggests the most balanced performance across all metrics. This balance is particularly crucial in glioma grading, where accurately distinguishing between LGG and GBM is vital for appropriate treatment planning. Both underestimating and overestimating the tumor grade can lead to either insufficient or overly aggressive treatment strategies, respectively.

Further analysis, through a direct comparison with prior work that uses the same dataset, particularly the study by Tasci et al., reveals the strengths of our method.[32] Our optimized LightGBM model, with an overall accuracy of 89.88 %, shows a marked improvement over the accuracy of 87.606 % reported by Tasci et al. This advancement underlines the effectiveness of our approach in glioma classification, demonstrating significant progress in reducing diagnostic errors. The improved accuracy of our model not only affirms its superiority in the context of existing methods but also highlights its potential to contribute to more reliable and effective treatment planning in neuro-oncology.

The ROC curve with a 95 % confidence interval (CI) depicted in Fig. 2 illustrates the comparative diagnostic performance of various ML models for glioma grading. The LightGBM model demonstrated the highest diagnostic accuracy, with an AUC of 0.95 (95 % CI: 0.92–0.98), highlighting its superior capability to differentiate between LGG and GBM cases. The Random Forest model closely followed with an AUC of 0.94 (95 % CI: 0.91–0.97), and the Support Vector Machine showed similarly strong discriminative power with an AUC of 0.93 (95 % CI: 0.89–0.96). The k-Nearest Neighbors model also performed robustly, achieving an AUC of 0.91 (95 % CI: 0.87–0.95). In comparison, the Naïve Bayesian and Decision Tree models had lower AUCs of 0.89 and 0.83, respectively. Specifically, the Decision Tree model's CI ranged from 0.77 to 0.89, indicating a relatively lower effectiveness in distinguishing glioma grades than other models.

The high AUC of 0.95 for the LightGBM model, along with its narrow confidence interval (0.92–0.98), underscores its reliability and potential as a valuable tool for clinicians. This model's strong ability to rank glioma grade probabilities accurately may support clinical decision-making, enabling more precise identification and classification of tumor types.

Table 2
Performance of ML models for glioma grading.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LightGBM	**89.88**	96.67	86.14	**91.10**
Random Forest	86.90	89.90	**88.12**	89.00
Naïve Bayesian	86.31	**98.75**	78.22	87.29
Decision Tree	86.31	89.00	**88.12**	88.56
Support Vector Machine	85.71	91.40	84.16	87.63
K-Nearest Neighbors	83.93	89.36	83.17	86.15

Note: Bold values indicate best results

4.2. Model Explanation with SHAP

To explain the importance of various features in the LightGBM model used for glioma grading, we visualized the SHAP bar plot presented in Fig. 3. The length of each bar represents the mean absolute SHAP value for each feature, indicating the average impact of that feature on the model's output. A longer bar denotes a higher impact on the model's predictions.

It can be observed that the IDH1 gene mutation has the highest mean SHAP value (+0.31), suggesting it is the most influential feature in the model for predicting glioma grade. This is consistent with medical research that recognizes IDH1 as a crucial biomarker in glioma, often differentiating LGG from GBM.

Age at diagnosis, with a mean SHAP value of + 0.09, is the second most influential feature. This supports the clinical understanding that age is a significant prognostic factor in glioma, with younger patients generally having better outcomes, possibly due to a higher likelihood of LGG or a better response to treatment.

The SHAP values for CIC and ATRX mutations at + 0.05 and + 0.04, respectively, suggest that while they have a smaller influence on glioma grading in the LightGBM model compared to the IDH1 mutation, they are still significant in differentiating tumor subtypes and have implications for patient outcomes. The TP53 mutation holds a modest yet notable impact with a SHAP value of + 0.03, indicative of its role in cancer biology and its specific relevance to glioma progression. The sum of the mean SHAP values for the 18 other features amounts to + 0.11, which, although each feature individually contributes less than IDH1 or Age at diagnosis, their collective weight is meaningful.

The bee swarm plot in Fig. 4 offers a detailed view of SHAP values in the LightGBM model for glioma grading, contrasting with the summary bar plot. It showcases individual feature contributions to predictions, revealing the variability and how specific feature values impact predictions. Pink and blue dots indicate high and low feature values, respectively, influencing predictions towards GBM or LGG, as depicted in the color legend. Pink dots contribute to pushing the model's prediction towards GBM, and blue dots push it towards LGG.

Upon closer examination of the bee swarm SHAP plot for glioma grading, it becomes evident that the positioning of the dots for IDH1 indicates that higher SHAP values (dots skewed to the right and likely colored pink) are associated with classifying LGG, while lower SHAP values (dots skewed to the left and likely colored blue) are associated with classifying GBM. This suggests that the model interprets higher values of IDH1 as indicative of LGG, which aligns with the medical understanding that IDH1 mutations are more common in LGG and lower values as indicative of GBM.[50]

The age at diagnosis shows a mix of pink and blue dots spread across both sides of the zero line. This distribution suggests that the age at diagnosis influences the model's predictions in a more complex manner. Younger ages contribute to the prediction of LGG, whereas older ages push the prediction towards GBM. The spread of the dots indicates variability in how age contributes to the prediction, with no clear skewness towards one class, which reflects the diverse nature of glioma incidence across ages.

For the feature CIC, the dots are concentrated around the zero line, with a somewhat even spread of pink and blue dots on both sides. This indicates that the presence or absence of CIC alterations has a variable impact on the model's prediction, not overwhelmingly pointing to either GBM or LGG. The color coding suggests that both high and low values of this feature are found across the dataset, and its influence on the prediction outcome is not as pronounced as some of the other features.

ATRX shows a pattern similar to CIC, with dots close to the zero line, implying that this feature has a nuanced effect on the model's

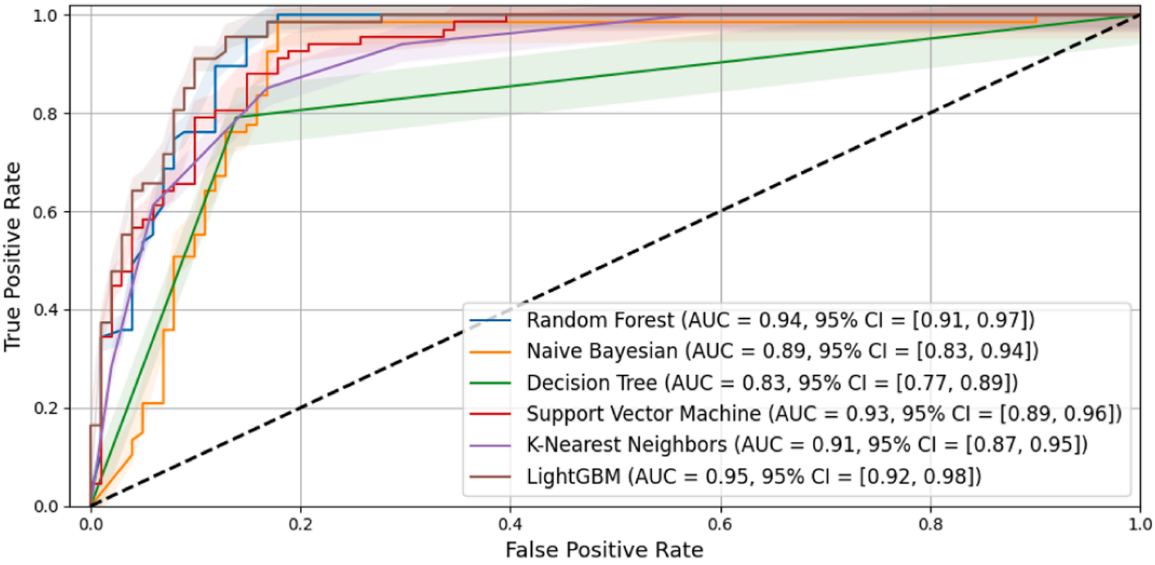

Fig. 2. ROC curve of ML models for glioma grading.

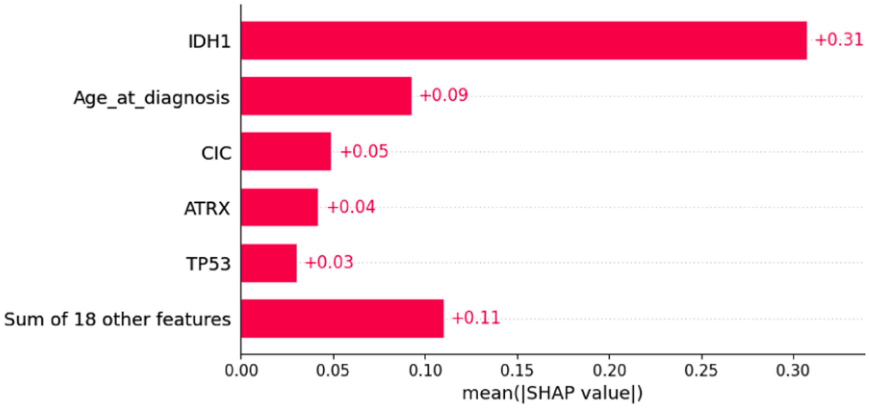

Fig. 3. SHAP bar plot.

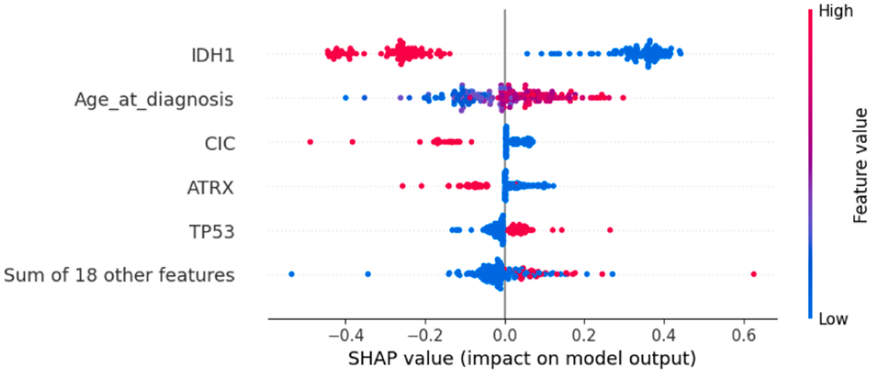

Fig. 4. SHAP bee swarm plot.

prediction. The SHAP values for ATRX do not show a strong bias toward either prediction class, which could imply that while ATRX mutations are a factor in glioma biology, they may not be a dominant feature in the grading distinction within this particular model or dataset.

TP53, on the other hand, shows a more distinct pattern, with a cluster of pink dots to the right. This suggests that higher values of TP53, which may correspond to mutations or increased expression, are influential in the model's prediction of GBM. The presence of fewer dots to the left suggests that lower values of TP53 are less influential or less common in predictions for LGG.

Finally, the sum of 18 other features row at the bottom indicates the collective impact of the remaining features not individually listed. The spread of dots across the SHAP value spectrum suggests that these features together contribute a variable influence on the model's predictions. This aggregation can obscure the impact of individual features within this group but indicates that combined, they do have a discernible impact on the model's output.

5. Discussion

The results of this study demonstrate the promising capabilities of integrating LightGBM with XAI techniques, specifically SHAP, in the grading of gliomas. This integration addresses a crucial need in neuro-oncology for accurate and interpretable ML models capable of handling the complexity of medical data for brain tumor classification.

The superior accuracy of LightGBM over traditional ML models, such as Random Forest, Naïve Bayesian, Decision Tree, Support Vector Machine, and k-Nearest Neighbors, underscores its efficacy in navigating the intricacies of glioma grading. This is a critical observation since the distinction between LGG and GBM directly impacts treatment strategies and patient outcomes. However, the study also indicates that no single model uniformly excels across all performance metrics, including precision and recall. This underscores the necessity of a balanced evaluation of models, where the trade-offs between sensitivity and specificity are carefully weighed, especially in clinical settings where the implications of model predictions are significant.[51]

A key strength of this study lies in the application of SHAP values for model explanation, which enhances transparency and clinician trust in the model's decision-making process. The identification of IDH1 gene mutation as a pivotal feature in glioma grading through SHAP values not only aligns with existing medical knowledge but also reinforces the model's clinical utility by providing insights into the biological underpinnings of its predictions.

Despite these strengths, the study acknowledges limitations that could influence its applicability and generalizability. The dataset used, while comprehensive, may not capture the full spectrum of glioma heterogeneity encountered in diverse clinical scenarios. This limitation poses a risk to the model's generalizability across different patient populations and glioma subtypes.[52] Moreover, the study's focus on specific genetic and clinical features, albeit insightful, may omit other critical factors contributing to glioma behavior and prognosis.

Future research directions should aim at broadening the dataset to include a more diverse array of glioma cases, thereby enhancing the model's robustness and applicability in real-world clinical settings. Expanding the scope of genetic and clinical features considered in the model could also provide a more nuanced understanding of glioma characteristics, potentially uncovering new predictors of tumor grade and behavior. Additionally, evaluating the model on other datasets is crucial to verify its generalizability and performance across different populations and settings. This approach will not only improve the diagnostic precision of glioma grading but also contribute to the development of personalized treatment plans, ultimately benefiting patient care in neuro-oncology.

6. Conclusion

This study's exploration into glioma grading using the LightGBM model has yielded significant insights, substantiating the model's robustness and accuracy. We have successfully demonstrated that through meticulous hyperparameter optimization and performance evaluation, the LightGBM model outshines traditional ML counterparts and underscores the model's capability to distinguish between LGG and GBM grades effectively. The application of SHAP values for model interpretation aligns with the principles of XAI, offering granular insights into the decision-making process and establishing the importance of features such as IDH1 mutations and patient age at diagnosis. However, the study acknowledges limitations, including the dataset's potential lack of representation for the full spectrum of glioma heterogeneity and the exclusion of certain genetic and clinical features, which may affect the model's generalizability. Future studies should address these limitations by expanding datasets and feature considerations and evaluating the model's performance across diverse populations. This approach promises to refine diagnostic accuracy and contribute to more personalized and effective treatment strategies, ultimately improving patient outcomes in neuro-oncology.

CRediT authorship contribution statement

Ghalieb Mutig Idroes: Writing – review & editing, Visualization, Methodology, Investigation. **Irsan Hardi:** Writing – original draft, Validation, Conceptualization. **Teuku Rizky Noviandy:** Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization.

Declaration of Competing Interest

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

Acknowledgments

The authors would like to express their sincere gratitude to Graha Primera Saintifika for their generous support to this study.

Data availability

The dataset used in this study was retrieved from the UCI Machine Learning Repository and is accessible at the following link: https://archive.ics.uci.edu/dataset/759/glioma+grading+clinical+and+mutation+features+dataset. The code utilized for this study can be accessed on GitHub at the following repository: https://github.com/trizkynoviandy/glioma_grading.

References

1. Chen R, Smith-Cohn M, Cohen AL, Colman H. Glioma subclassifications and their clinical significance. *Neurotherapeutics*. 2017;14(2):284–297. https://doi.org/10.1007/s13311-017-0519-x.
2. Wu W, Klockow JL, Zhang M, et al. Glioblastoma multiforme (GBM): an overview of current therapies and mechanisms of resistance. *Pharm Res*. 2021;171, 105780. https://doi.org/10.1016/j.phrs.2021.105780.
3. Kumthekar P, Raizer J, Singh S. Low-Grade Glioma. *In*. 2015:75–87. https://doi.org/10.1007/978-3-319-12048-5_5.
4. Suárez-García JG, Hernández-López JM, Moreno-Barbosa E, de Celis-Alonso B. A simple model for glioma grading based on texture analysis applied to conventional brain MRI. In: Sherman JH, ed. PLoS One. 15. 2020, e0228972. https://doi.org/10.1371/journal.pone.0228972.
5. Visser M, Müller DMJ, van Duijn RJM, et al. Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage Clin*. 2019;22, 101727. https://doi.org/10.1016/j.nicl.2019.101727.
6. El Naqa I, Karolak A, Luo Y, et al. Translation of AI into oncology clinical practice. *Oncogene*. 2023;42(42):3089–3097. https://doi.org/10.1038/s41388-023-02826-z.
7. Noviandy T.R., Alfanshury M.H., Abidin T.F., Riza H. Enhancing Glioma Grading Performance: A Comparative Study on Feature Selection Techniques and Ensemble Machine Learning. In: *2023 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*. IEEE; 2023:406-411. doi:10.1109/IC3INA60834.2023.10285778.
8. Idroes R, Noviandy TR, Maulana A, et al. Application of genetic algorithm-multiple linear regression and artificial neural network determinations for prediction of kovats retention index. *Int Rev Model Simul*. 2021;14(2):137. https://doi.org/10.15866/iremos.v14i2.20460.
9. Agustia M., Noviandy T.R., Maulana A., et al. Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances. In: *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*. IEEE; 2022:13-18. doi:10.1109/ICELTICs56128.2022.9932124.
10. Zheng Q, Zhao P, Zhang D, Wang H. MR-DCAE: Manifold regularization-based deep convolutional autoencoder for unauthorized broadcasting identification. *Int J Intell Syst*. 2021;36(12):7204–7238. https://doi.org/10.1002/int.22586.
11. Barzegar Behrooz A, Latifi-Navid H, da Silva Rosa SC, et al. Integrating multi-omics analysis for enhanced diagnosis and treatment of glioblastoma: a comprehensive data-driven approach. *Cancers (Basel)*. 2023;15(12):3158. https://doi.org/10.3390/cancers15123158.
12. Noviandy TR, Maulana A, Idroes GM, et al. Integrating genetic algorithm and LightGBM for QSAR modeling of acetylcholinesterase inhibitors in Alzheimer's Disease drug discovery. *Malacca Pharm*. 2023;1(2):48–54. https://doi.org/10.60084/mp.v1i2.60.

13. Rufo DD, Debelee TG, Ibenthal A, Negera WG. Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*. 2021;11(9):1714. https://doi.org/10.3390/diagnostics11091714.
14. Noviandy T.R., Maulana A., Idroes G.M., Irvanizam I., Subianto M., Idroes R. QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction. In: *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*. IEEE; 2023:220-225. doi:10.1109/COSITE60233.2023.10250039.
15. Nurdin Z., Hidayat T., Irvanizam I. Performance Comparison of Hybrid CNN-XGBoost and CNN-LightGBM Methods in Pneumonia Detection. In: *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*. IEEE; 2022:31-36. doi:10.1109/ICELTICs56128.2022.9932129.
16. Dwivedi R, Dave D, Naik H, et al. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput Surv*. 2023;55(9):1–33. https://doi.org/10.1145/3561048.
17. Noviandy T.R., Maulana A., Khowarizmi F., Muchtar K. Effect of CLAHE-based Enhancement on Bean Leaf Disease Classification through Explainable AI. In: *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*. IEEE; 2023:515-516. doi:10.1109/GCCE59613.2023.10315394.
18. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. 2021;32(11):4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314.
19. Muddamsetty SM, Jahromi MNS, Moeslund TB. Expert level evaluations for explainable AI (XAI) methods in the medical domain. *In*. 2021:35–46. https://doi.org/10.1007/978-3-030-68796-0_3.
20. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*. 2022;12(2):237. https://doi.org/10.3390/diagnostics12020237.
21. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners – Beyond saliency-based XAI approaches. *Eur J Radio*. 2023; 162, 110786. https://doi.org/10.1016/j.ejrad.2023.110786.
22. Antoniadi AM, Du Y, Guendouz Y, et al. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Appl Sci*. 2021;11(11):5088. https://doi.org/10.3390/app11115088.
23. Hou F., Cheng Z., Kang L., Zheng W. Prediction of Gestational Diabetes Based on LightGBM. In: *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*. ACM; 2020:161-165. doi:10.1145/3433996.3434025.
24. Noviandy TR, Nainggolan SI, Raihan R, Firmansyah I, Idroes R. Maternal health risk detection using light gradient boosting machine approach. *Info J Data Sci*. 2023;1(2): 48–55. https://doi.org/10.60084/ijds.v1i2.123.
25. Yang H, Chen Z, Yang H, Tian M. Predicting coronary heart disease using an improved LightGBM model: performance analysis and comparison. *IEEE Access*. 2023;11:23366–23380. https://doi.org/10.1109/ACCESS.2023.3253885.
26. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Inf Fusion*. 2022;81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011.
27. Carlens H. State of Competitive Machine Learning in 2022. *ML Contests Res*. Published online 2023.
28. McElfresh D, Khandagale S, Valverde J, et al. When do neural nets outperform boosted trees on tabular data? *Adv Neural Inf Process Syst*. 2024;36.
29. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
30. Shboul ZA, Chen J, M. Iftekharuddin K. Prediction of molecular mutations in diffuse low-grade gliomas using MR imaging features. *Sci Rep*. 2020;10(1):3711. https://doi.org/10.1038/s41598-020-60550-0.
31. Wu S, Meng J, Yu Q, Li P, Fu S. Radiomics-based machine learning methods for isocitrate dehydrogenase genotype prediction of diffuse gliomas. *J Cancer Res Clin Oncol*. 2019;145(3):543–550. https://doi.org/10.1007/s00432-018-2787-1.
32. Tasci E, Zhuge Y, Kaur H, Camphausen K, Krauze AV. Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. *Int J Mol Sci*. 2022;23(22):14155. https://doi.org/10.3390/ijms232214155.
33. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.
34. Noviandy TR, Maulana A, Idroes GM, et al. Ensemble machine learning approach for quantitative structure activity relationship based drug discovery: a review. *Info J Data Sci*. 2023;1(1):32–41. https://doi.org/10.60084/ijds.v1i1.91.
35. Jaiyeoba O, Ogbuju E, Yomi OT, Oladipo F. Development of a model to classify skin diseases using stacking ensemble machine learning techniques. *J Comput Theor Appl*. 2024;2(1):22–38. https://doi.org/10.62411/jcta.10488.
36. Breslin W, Pham D. Machine learning and drug discovery for neglected tropical diseases. *BMC Bioinforma*. 2023;24(1):165. https://doi.org/10.1186/s12859-022-05076-0.
37. Noviandy TR, Idroes GM, Hardi I. Machine learning approach to predict AXL kinase inhibitor activity for cancer drug discovery using xgboost and bayesian optimization. *J Soft Comput Data Min*. 2024;5(1):46–56.
38. Maulana A, Faisal FR, Noviandy TR, et al. Machine learning approach for diabetes detection using fine-tuned XGBoost algorithm. *Info J Data Sci*. 2023;1(1):1–7. https://doi.org/10.60084/ijds.v1i1.72.
39. Suhendra R, Suryadi S, Husdayanti N, et al. Evaluation of gradient boosted classifier in atopic dermatitis severity score classification. *Heca J Appl Sci*. 2023;1(2):54–61. https://doi.org/10.60084/hjas.v1i2.85.
40. Zheng Q, Saponara S, Tian X, Yu Z, Elhanashi A, Yu R. A real-time constellation image classification method of wireless communication signals based on the lightweight network MobileViT. *Cogn Neurodyn*. 2024;18(2):659–671. https://doi.org/10.1007/s11571-023-10015-7.
41. Zheng Q, Tian X, Yu Z, et al. MobileRaT: a lightweight radio transformer method for automatic modulation classification in drone communication systems. *Drones*. 2023; 7(10):596. https://doi.org/10.3390/drones7100596.
42. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13(2).
43. Idroes R, Noviandy TR, Maulana A, Suhendra R, Sasmita NR. ANFIS-based QSRR modelling for kovats retention index prediction in gas chromatography. *Info J Data Sci*. 2023;1(1):1–14. https://doi.org/10.60084/ijds.v1i1.73.
44. Zheng Q, Wang R, Tian X, et al. A real-time transformer discharge pattern recognition method based on CNN-LSTM driven by few-shot learning. *Electr Power Syst Res*. 2023;219, 109241. https://doi.org/10.1016/j.epsr.2023.109241.
45. Zheng Q, Tian X, Yang M, Wu Y, Su H. PAC-Bayesian framework based drop-path method for 2D discriminative convolutional network pruning. *Multidimens Syst Signal Process*. 2020;31(3):793–827. https://doi.org/10.1007/s11045-019-00686-z.
46. Idroes GM, Maulana A, Suhendra R, et al. TeutongNet: a fine-tuned deep learning model for improved forest fire detection. *Leuser J Environ Stud*. 2023;1(1):1–8. https://doi.org/10.60084/ljes.v1i1.42.
47. Noviandy TR, Maulana A, Emran TB, Idroes GM, Idroes R. QSAR classification of beta-secretase 1 inhibitor activity in Alzheimer's Disease using ensemble machine learning algorithms. *Heca J Appl Sci*. 2023;1(1):1–7. https://doi.org/10.60084/hjas.v1i1.12.
48. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.
49. Le TTH, Kim H, Kang H, Kim H. Classification and explanation for intrusion detection system based on ensemble trees and SHAP method. *Sensors*. 2022;22(3): 1154. https://doi.org/10.3390/s22031154.
50. Dono A, Ballester LY, Primdahl D, Esquenazi Y, Bhatia A. IDH-mutant low-grade glioma: advances in molecular diagnosis, management, and future directions. *Curr Oncol Rep*. 2021;23(2):20. https://doi.org/10.1007/s11912-020-01006-6.
51. Varoquaux G., Colliot O. Evaluating Machine Learning Models and Their Diagnostic Value. In:; 2023:601-630. doi:10.1007/978-1-0716-3195-9_20.
52. Yom SS, Deville C, Boerma M, Carlson D, Jabbour SK, Braverman L. Evaluating the generalizability and reproducibility of scientific research. *Int J Radiat Oncol*. 2022; 113(1):1–4. https://doi.org/10.1016/j.ijrobp.2022.02.002.