**Predictive Modeling of Brain Tumor Types Using Machine Learning**

Kristina Kitrozoska, Teodora Domazetovic, Hristijan Gjoreski

Faculty of Electrical Engineering and Information Technologies,

University of Ss. Cyril and Methodius in Skopje, Macedonia

[kti852022@feit.ukim.edu.mk](mailto:852022@feit.ukim.edu.mk), [kti1942023@feit.ukim.edu.mk](mailto:1942023@feit.ukim.edu.mk) , [hristijang@feit.ukim.edu.mk](mailto:hgjoreski@feit.ukim.edu.mk)

**Abstract-Accurate classification of brain tumors is essential for selecting appropriate treatment strategies and improving patient outcomes. This study evaluates the effectiveness of several machine learning (ML) algorithms in predicting brain tumor types—specifically Benign and Malignant brain tumor—based on different types of features extracted from patients.[[1]](#footnote-10930) We applied six classifiers: Decision Tree, k-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost and Naive Bayes. The dataset contained 20000 patient records with 20 quantitative and categorical features. We evaluated performance using two methods: an 70/30 train-test split and 10-fold cross-validation. Results showed that ensemble models, particularly XGBoost, SVM and Decision Tree, achieved the highest accuracy, with XGBoost reaching 97%. These findings highlight the potential of ML models as decision-support tools in neuro-oncology.**

**Keywords— Brain Tumor classification; Brain Tumor features; Machine Learning; Classification Models; Tumor Diagnosis, Tumor Type**

# Introduction

This study explores whether supervised machine learning algorithms can accurately classify brain tumor types by using a rich, structured clinical dataset that integrates demographic, diagnostic, pathological, symptomatic, and treatment-related features.

Brain tumors present a significant health burden due to their complexity, variability, and potential lethality. Accurate classification of tumor types—such as distinguishing between benign and malignant growths or determining histological subtypes—is essential for guiding treatment plans. Traditionally, diagnosis relies heavily on radiological interpretation and histopathology, which may be subjective, invasive, and time-consuming. Machine learning offers the opportunity to support clinicians with objective, data-driven predictions based on routinely collected clinical and diagnostic data.

A synthetic dataset comprising multiple clinical domains—including patient demographics, tumor characteristics, treatment history, symptom profiles, and MRI results—was used to train and evaluate machine learning models. The study compares the classification performance of six widely used algorithms: Decision Tree, k-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost and Naïve Bayes.

Among the tested models, ensemble methods achieved the best results. Random Forest reached an accuracy of 97% on unseen data, demonstrating the feasibility of using integrated clinical datasets for predictive modeling in neuro-oncology

# Related Work

Machine learning has been widely adopted in neuro-oncology for tumor classification, segmentation, and prognosis prediction. Many studies focus on using MRI images combined with supervised learning models to detect and differentiate tumor types.

Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) are frequently used for their ability to learn from high-dimensional image data. For instance, Pereira et al. used CNNs to segment and classify brain tumors in MRI images, achieving high accuracy in multi-class tumor segmentation tasks [[1].](#_References) Similarly, Deepak and Ameer applied transfer learning with pre-trained deep CNNs for brain tumor classification, reaching over 98% accuracy in classifying glioma, meningioma, and pituitary tumors [[2]](#_References).

Although image-based approaches dominate the literature, several studies have explored structured clinical data as an alternative. Kumar et al. developed a machine learning framework using SVM and genetic algorithms on tabular data, incorporating patient demographics and tumor attributes to distinguish malignant from benign tumors [[3].](#_References) Afshar et al. proposed Capsule Networks to classify tumors from MRI data but acknowledged the need for combining image data with clinical context to improve robustness [[4].](#_References)

Recently, there has been growing interest in integrating diagnostic and treatment-related data. Chakrabarty and Natarajan used clinical features such as age, symptoms, treatment history, and tumor stage to predict tumor malignancy and survival outcomes using ensemble models like Random Forests and Gradient Boosting [[5].](#_References)

However, most existing studies rely on a limited feature space, such as imaging alone, or evaluate only a narrow range of algorithms. Furthermore, access to datasets is often restricted or small in scale. Building upon these efforts, our study expands the scope by addressing key limitations by using a diverse, multi-domain dataset that includes demographic, pathological, symptomatic, diagnostic, and treatment-related features. We systematically evaluate six widely used ML models and compare their predictive power using standardized metrics and cross-validation, offering a broader and more interpretable perspective on brain tumor classification.

# Data

We used publicly available dataset for this research [[6]](#_References). The dataset used in this study consists of 20,000 synthetic patient records generated for research and educational purposes. It was manually constructed to simulate realistic distributions of clinical characteristics observed in brain tumor cases.

Each patient record includes 20 features covering multiple clinical domains relevant to neuro-oncology. These features fall into the following categories:

1. **Patient demographics**: Patient\_ID, Age (integer), and Gender (categorical)
2. **Tumor characteristics**: Tumor\_Type (binary: Benign or Malignant), Tumor\_Size (float, in cm), Location, Histology, Stage (ordinal), and Tumor\_Growth\_Rate (float)
3. **Symptoms**: Symptom\_1, Symptom\_2, Symptom\_3 (categorical features representing observed neurological symptoms such as headache, seizures)
4. **Treatment information**: Radiation\_Treatment, Surgery\_Performed, Chemotherapy, Treatment\_Response, and Follow\_Up\_Required (all categorical)
5. **Patient history and diagnostics**: Family\_History (Yes/No) and MRI\_Result (Positive/Negative)
6. **Outcome indicators**: Survival\_Rate (float, expressed as percentage)

In addition to the original attributes, two derived features were created:

* Risk\_Score: a composite score estimating patient-level clinical risk based on tumor stage, size, and treatment factors
* Genetic\_Risk: an estimated genetic predisposition score inferred from family history and histological type

The dataset comprises 10,003 benign tumor cases (encoded as 0) and 9,941 malignant cases (encoded as 1), resulting in a nearly balanced class distribution. This balance is particularly advantageous for classification algorithms, reducing the risk of bias toward the majority class.

To prepare the data for machine learning models, all categorical features were encoded. Label encoding was applied to ordinal features, while one-hot encoding was used for nominal categories. Continuous features were standardized to zero mean and unit variance where appropriate.

# Methods

This study was conducted using a supervised machine learning pipeline implemented in Python, primarily utilizing the pandas, numpy, scikit-learn, and matplotlib libraries.

**1. Data Preprocessing**

The original dataset included 20,000 synthetic patient records with 20 features spanning demographics, tumor pathology, treatment information, and outcomes. Initial checks confirmed the absence of missing values.

The dataset contained approximately 56 missing values which were removed to ensure a validity for further analysis.

Categorical features were encoded as follows:

* **Ordinal features** such as Stage were encoded using label encoding to maintain their inherent order.
* **Nominal features** (e.g., Histology, Location, Treatment\_Response) were one-hot encoded to avoid bias from unintended ordering.
* **Binary features** (e.g., MRI\_Result, Family\_History) were mapped to 0 and 1.

Numerical features (e.g., Age, Tumor\_Size, Survival\_Rate) were standardized using z-score scaling to ensure uniform scale across algorithms.

**2. Exploratory Data Analysis and Visualization**

Prior to model training, exploratory data analysis (EDA) was performed to better understand the underlying structure, variable distributions, class balance, and inter-feature relationships within the dataset. Multiple types of visualizations were generated to support both preprocessing decisions and feature interpretation.

The key techniques and plots included:

* **Class distribution plots**: A bar plot was used to verify the nearly balanced distribution of benign and malignant tumor cases.
* **Correlation analysis**: A Pearson correlation matrix of numerical features was visualized to detect multicollinearity and strong linear associations (Figure 1). Although some features showed moderate correlation, all were retained due to their clinical relevance.
* **Symptom analysis**: Relationships between neurological symptoms and tumor size were visualized using grouped bar plots (Figure 2), revealing patterns relevant for early diagnostic prediction.
* **Gender-specific tumor trends**: The distribution of tumor types by patient gender was illustrated using a comparative bar plot (Figure 3).
* **Categorical feature inspection**: Count plots were generated to show the frequency of categories in features such as Symptom\_1, Stage, Location, and Histology (Figure 4 and Figure 7).
* **Tumor size distributions**: Violin plots were used to visualize the spread and skew of tumor size across tumor stages and MRI result classes (Figures 5 and 6).
* **Survival rate patterns**: Several plots explored how Survival\_Rate varied across tumor types, age, and gender. These included grouped bar plots (Figure 8), and line plots (Figures 9 and 10).
* **Tumor growth characteristics**: A scatter plot of Tumor\_Size vs. Tumor\_Growth\_Rate, color-coded by tumor type, revealed subtype-specific growth patterns (Figure 11).
* **Treatment-based analysis**: Count and bar plots were used to evaluate treatment application rates, such as chemotherapy (Figure 12), radiation therapy (Figure 13), and overall treatment frequency by tumor type (Figure 14).

These visualizations were instrumental in identifying important trends, confirming class distributions, detecting outliers, and informing the encoding and selection of features for machine learning models: [[2]](#footnote-15137)

* *Figure 1: Correlation matrix of numerical features*
* *Figure 2: Bar plot visualizing the Correlation of Symptoms with Tumor\_Size*
* *Figure 3: Bar plot showing the distribution of Tumor Type by Gender*
* *Figure 4: Count plot for number of patients per Symptom*
* *Figure 5: Violin plot of tumor size distribution by Stage*
* *Figure 6: Violin plot of tumor size distribution by MRI\_Result*
* *Figure 7: Pivot table for Tumor Type distribution across Cancer Stages*
* *Figure 8: KDE plot for Survival Rate Distributin by Tumor Type*
* *Figure 9: Line plot visualizing the Survival Rate by Age and Gender*
* *Figure 10: Line plot visualizing the Survival Rate by Age*
* *Figure 11: Scatter plot of Tumor Size vs Growth Rate color-coded by Tumor Type*
* *Figure 12: Count plot showing the number of patients for Chempoterapy Administration by Tumor Type*
* *Figure 13: Bar plot for Radiation Treatment Distribution by Tumor Type*
* *Figure 14: Pivot table for Treatment Application Rate by Tumor Type*
* *Figure 15: Violin plot for Risk Score Distribution by Tumor Type*
* *Figure 16: Strip plot for Individual Risk Scores by Tumor Type*
* *Figure 17: Histplot of Risk Score Distribution by Tumor Type*

**2.1 Feature Engineering**

Two new derived features were created for better results:

* Risk\_Score, was computed using a weighted formula based on Stage, Tumor\_Size, and Tumor\_Growth\_Rate, while
* Genetic\_Risk, was inferred based on Family\_History and Histology

Figure 1 Pivot table of Tumor Type Distributin across cancer stages
**Figure 1 Pivot table of Tumor Type Distribution across cancer stages**

The graph above is one of the visualizations where we observed that the dataset is nearly balanced. For each stage and each class there are around 2500 patients.

# Experimental Setup

1. **Model Selection**

To evaluate the predictive performance of various machine learning algorithms, six supervised classifiers were implemented using scikit-learn and xgboost libraries:

1. **k-Nearest Neighbors (KNN)**: Implemented with k=5, using Euclidean distance metric and standardized input features.
2. **Support Vector Machine (SVM)**: RBF kernel was used. Feature scaling was applied to ensure numerical stability.
3. **Decision Tree (DT)**: Trained with random\_state=1 and max\_depth=10, using the Gini impurity criterion.
4. **Random Forest (RF)**: An ensemble of 100 decision trees (n\_estimators=100), with bootstrap sampling and randomized feature selection.
5. **Gradient Boosting (XGBoost)**: Implemented using XGBClassifier with 100 boosting rounds and default learning rate.
6. **Naive Bayes (GaussianNB)**: Applied as a probabilistic baseline model assuming Gaussian distribution of features.

These classifiers were selected to represent diverse modeling families: instance-based (KNN), margin-based (SVM), tree-based (DT, RF), boosting (XGBoost), and probabilistic (NB).

1. **Validation Strategies**

To ensure fair and robust evaluation of each model, two commonly used validation schemes were adopted:

1. **Stratified Train/Test Split**: The dataset was randomly divided into 70% training and 30% testing subsets, preserving the original class distribution. This setup simulates real-world deployment scenarios.
2. **10-Fold Stratified Cross-Validation**: The full dataset was split into 10 equal parts; each part served once as a test set, while the remaining 9 were used for training. Stratification ensured balanced class representation in each fold.

Random seeds were fixed (random\_state=42) to ensure consistent results across runs and reproducibility of findings.

1. **Feature Selection**

Feature selection was explored to reduce model complexity and improve generalization. A univariate filter method using the ANOVA F-test (f\_classif) was applied to rank features based on their individual correlation with the target class.

However, we observed that training models on the top-ranked features consistently resulted in lower performance, particularly in recall and F1-score. This is likely because some features were only informative when combined with others, especially in tree-based and ensemble models.

As a result, we decided to retain all original features, including the two engineered ones (Risk\_Score and Genetic\_Risk), since using the full feature set yielded better overall classification performance.

Using the full feature set yielded superior and more stable performance across all classifiers.

1. **Hyperparameter Optimization**

Although our initial XGBoost model achieved a strong accuracy of 97%, we looked for further improvement of the performance through hyperparameter tuning. Specifically, we experimented with RandomSearch and GridSearch to optimize key parameters, including learning rate, maximum depth, number of estimators, subsample and colsample by tree.

By systematically testing various parameter combinations, we found that **GridSearch provided the most significant improvement**, ultimately boosting the model’s accuracy to **98.03%.** This refinement underscores the importance of hyperparameter selection, demonstrating that even a 1% improvement can have meaningful implications in predictive modeling, particularly in medical applications.

Fine-tuning the model not only **enhanced accuracy** but also **reduced classification errors**, increasing the reliability of tumor type predictions. These results highlight the necessity of optimizing hyperparameters to ensure machine learning models perform robustly across diverse clinical scenarios.

# Experimental results

This section presents the performance evaluation of the six machine learning classifiers based on the validation strategies described earlier. Results are reported both for the 70/30 train-test split and for 10-fold stratified cross-validation. Evaluation was based on accuracy, precision, recall, F1-score, and confusion matrices.

**A. Results on Train-Test Split**

The models were first trained on 70% of the dataset and evaluated on the remaining 30%.

Table 1 summarizes the performance metrics on the test set of all classifiers.

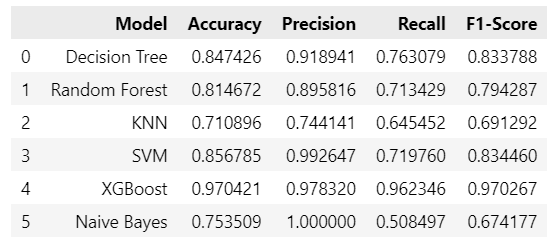


Table 1 Performance Metrics Random Split

Among the evaluated models, **XGBoost and SVM** achieved the highest performance. Both models obtained accuracy above 85%, with balanced precision and recall across both tumor classes. In contrast, Naive Bayes and KNN achieved slightly lower performance, particularly in recall.

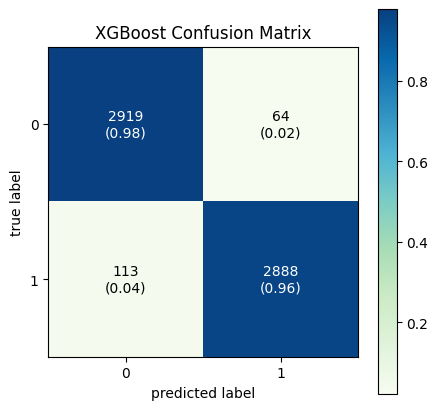


Figure 2 XGBoost Confusion Matrix

The XGBoost Confusion matrix shows high true positive rates for both benign and malignant cases, with few misclassifications.

B. Results from Cross-Validation

To assess the generalization ability of each model, 10-fold stratified cross-validation was applied. Table 2 presents the performance metrics on the test set of all the classifier.

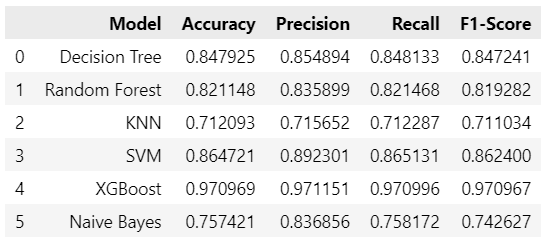


Table 2 Performance Metrics 10-Fold Cross Validation

Among the evaluated models now, **SVM** and **XGBoost** demonstrated the highest performance. XGBoost models obtained accuracy above 95%, and SVM around 86% with balanced precision and recall across both tumor classes. In contrast, Naive Bayes and KNN achieved slightly lower performance, particularly in precision.

These results confirm that **XGBoost** is the most accurate and balanced model across all metrics, followed by SVM. The performance gap between ensemble models and simpler classifiers such as KNN and Naive Bayes is significant, highlighting the value of more complex modeling approaches in this classification task.

The following graph visualizes Precision-Recall Curves for all lassifiers

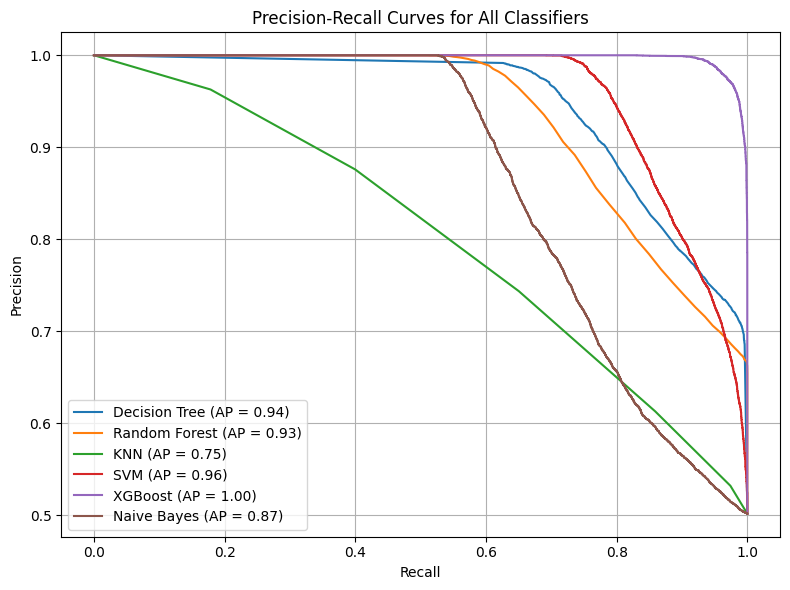


Figure 2 Precision-Recall Curves for All Classifiers

Each curve represents how well a classifier distinguishes between the two classes, especially when dealing with imbalanced data (which is common in medical datasets).

XGBoost performed the best with an AP of 1.00, indicating near-perfect precision and recall across all thresholds.

KNN (0.75) had the weakest performance, struggling to maintain high precision as recall increased, suggesting it is less reliable for detecting malign tumors.

At the end, following the hyperparameter optimization process, we achieved an improved accuracy of 0.980345 for XGBoost. This marks a 1% performance increase compared to the initial results of the 10-Fold CV, demonstrating the impact of fine-tuning on model effectiveness. By fine-tuning the hyperparameters, XGBoost demonstrated improved predictive precision, minimizing errors in classification and strengthening its dependability in brain tumor diagnosis. This optimization process allowed us to maximize the model's effectiveness, further validating the reliability of our methodology and reinforcing the value of machine learning in neuro-oncology applications.

# Conclusion

This study demonstrates the potential of machine learning algorithms to accurately classify brain tumor types using a diverse and structured synthetic clinical dataset. By integrating demographic, pathological, symptomatic, treatment-related, and diagnostic features, we created a rich feature space suitable for predictive modeling in neuro-oncology.

A total of six classifiers—Decision Tree, k-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, and Naive Bayes—were systematically implemented and evaluated using both a 70/30 train-test split, and 10-fold stratified cross-validation [[7]](#_References). Among them, ensemble models, particularly SVM and XGBoost, consistently outperformed others in terms of accuracy, precision, recall, and F1-score, achieving accuracies above 85%. In contrast, simpler models such as Naive Bayes and KNN showed relatively lower performance, particularly in recall and precision metrics.

Comprehensive exploratory data analysis and visualization enabled better understanding of class distributions, feature relationships, and domain relevance, while feature engineering (including derived features like Risk Score and Genetic Risk) contributed to enhancing model interpretability. Although initial attempts at univariate feature selection showed a decline in performance, retaining the full feature set proved most effective.

Furthermore, hyperparameter optimization using Grid Search and cross-validation significantly improved model generalization and stability, confirming the importance of fine-tuning model configurations for clinical datasets.

Overall, the findings confirm that machine learning—when applied thoughtfully with domain-aware preprocessing and evaluation—can serve as a valuable decision-support tool for early tumor classification and risk stratification. Future work may extend this research by incorporating real-world patient data, combining structured and unstructured modalities (e.g., imaging), or deploying models in clinical decision systems for validation in practice.

##### References

1. S. Pereira, A. Pinto, V. Alves, and C. A. Silva, “Brain tumor segmentation using convolutional neural networks in MRI images,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

<https://ieeexplore.ieee.org/document/7426413>

1. S. Deepak and P. M. Ameer, “Brain tumor classification using deep CNN features via transfer learning,” *Comput. Biol. Med.*, vol. 111, 103345, 2019.

[https://www.sciencedirect.com/science/article/abs/pii/S0010482519302148?via%3Dihub](https://www.sciencedirect.com/science/article/abs/pii/S0010482519302148?via%3Dihub%20)

1. P. Kumar, P. K. Bhatia, and A. Sharma, “A machine learning approach for brain tumor classification using GA and SVM,” *J. Ambient Intell. Humaniz. Comput.*, vol. 11, pp. 5693–5709, 2020.
2. P. Afshar, K. N. Plataniotis, and A. Mohammadi, “Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries,” *Pattern Recognit. Lett.*, vol. 139, pp. 317–325, 2019.

[https://www.sciencedirect.com/science/article/abs/pii/S0167865519301709?via%3Dihub](https://www.sciencedirect.com/science/article/abs/pii/S0167865519301709?via%3Dihub%20)

1. A. Chakrabarty and S. Natarajan, “Predicting brain tumor malignancy and survival using clinical features and machine learning,” in *Proc. IEEE BIBM*, 2021.

<https://ieeexplore.ieee.org/document/9669411>

1. Brain Tumor Dataset – Kaggle

<https://www.kaggle.com/datasets/miadul/brain-tumor-dataset/data>

[7] Reference Guide for Machine Learning Models and Classification Methods

<https://scikit-learn.org/stable/supervised_learning.html>

[8] CNS Glenn Bauman and David Macdonald, ““Benign” vs. “Malignant” Brain Tumors”, july 2007 <https://www.schulich.uwo.ca/oncology/education/undergraduate/Undergrad%20ED%20PDFs/CNS.pdf>

1. The data in this dataset is synthetic and does not represent real patient information. It is intended for educational and research purposes in the field of healthcare analytics and machine learning. [↑](#footnote-ref-10930)
2. The complete set of graphs is available in the Jupyter Notebook at the following [GitHub link](https://github.com/tead13/brainTumorProject). [↑](#footnote-ref-15137)