Summary Report - Glioma Grading using ML

Group 5 – ML2 Group Project

Introduction

This project aimed to develop a machine learning classification model capable of distinguishing between Low-Grade Glioma (LGG) and Glioblastoma Multiforme (GBM) diseases, using patients' clinical and mutation features.

Early identification of glioma grade is essential in oncology since GBM is the most aggressive and lethal subtype, requiring immediate and intensive treatment, while LGG often follows a slower progression and allows for more conservative management strategies.

The dataset used originates from Kaggle, it was adapted from The Cancer Genome Atlas (TCGA). It contains a mixture of clinical, demographic, and genomic variables.

The goal was to build an accurate and interpretable classifier that could support clinicians in predicting glioma grade based on non-imaging molecular and clinical data.

Two powerful ensemble methods, Random Forest (bagging-based) and XGBoost (boosting-based), were trained, tuned and compared through GridSearchCV with Stratified K-Fold Cross-Validation and Recall.

1. Technical Approach and Methodology

The dataset contains roughly 1,500 patient records and over 30 features that combine clinical, demographics, and genetics information, including patients' age, gender, race, diagnostic category, and mutation indicators for key genes.

The target variable is Grade, encoded as:

- 0 = Low-Grade Glioma (LGG)
- 1 = Glioblastoma Multiforme (GBM)

The key variables in the dataset are the following:

- Clinical and Demographic: Age_at_diagnosis, Gender, Race, Primary_Diagnosis
- Genetic Mutation Indicators: Binary flags (MUTATED/NOT_MUTATED) for genes such as IDH1, TP53, ATRX, EGFR, PTEN, CIC, FUBP1, NOTCH1, NF1, PIK3CA among others.

An initial exploratory analysis revealed that:

- Age: GBM patients are significantly older than LGG patients (mean ≈ 58.9 vs 46.2 years).
- Diagnosis type: "Glioblastoma" was almost exclusively associated with grade 1 (GBM), while "Astrocytoma" and "Oligodendroglioma" were typically grade 0 (LGG).
- Mutation patterns:
  - IDH1 mutation was dominant in LGG patients (present in ~80% of LGG cases).
  - EGFR and PTEN mutations were highly prevalent among GBM patients.
  - TP53 and ATRX mutations were informative across both categories but with different co-occurrence patterns.

These variables provided a strong biological rationale for their inclusion in the classification model.

2. Data Preprocessing

To prepare the dataset for machine learning algorithms, several cleaning and encoding steps were performed:

- Missing Values: Non-informative placeholders such as '--' and 'not reported' were replaced with 'Unknown' to maintain record completeness without bias.
- Age transformation: The variable Age_at_diagnosis originally contained textual formats such as "55 years 210 days." A custom parser was applied to extract numeric ages (in years), followed by binning into five age ranges: 0–20, 21–40, 41–60, 61–80, and 80+ years.
  These age bins were later one-hot encoded and contributed to model interpretability.
- Categorical Encoding: Categorical Variables (Gender, Race, Primary_Diagnosis, Age_bin) were One-Hot Encoded using scikit-learn's OneHotEncoder(drop='first') to avoid multicollinearity.
- Mutation Features: Mutation columns were binary-encoded (MUTATED = 1, NOT_MUTATED = 0).
- Feature Correlation and Redundancy: Two thresholds for feature reduction, if features had a redundancy ≥ 0.99 or relevance ≤ 0.0001, they were discarded.

All in all, the final dataset contained only encoded numeric features, suitable for tree-based models.

3. Model Selection Rationale

Two ensemble learning algorithms have been compared:

Model	Principle	Strengths
Random Forest	Bagging (Bootstrap Aggregation)	Robust to noise, reduces overfitting, easy to interpret
XGBoost	Gradient Boosting	Sequential error correction, high accuracy, handles class imbalance well

Both models perform well with tabular data and categorical encodings.
The choice of Recall as the scoring metric reflected the clinical goal: maximizing thresholds, not just fixed accuracy.

4. Hyperparameter Tuning and Optimization

Each model was then tuned with GridSearchCV (systematic search over defined hyperparameter grids) combined with Stratified K-Folds (k=3) cross-validation.
This method ensures fair evaluation of each parameter combination, while preserving class balance in every fold.

The parameters tested in Random Forest are:
n_estimators: [50, 100, 200]
max_depth: [3, 5, 7]
min_samples_split: [10, 20]
min_samples_leaf: [5, 10]
max_features: ['sqrt', 'log2']

The parameters tested in XGBoost are:
n_estimators: [50, 100, 200]
max_depth: [3, 5]
learning_rate: [0.01, 0.05, 0.1]
subsample: [0.7, 0.8, 1.0]
colsample_bytree: [0.6, 0.8, 1.0]
reg_alpha: [0, 0.1, 1.0]
reg_lambda: [1.0, 5.0]

5. Results and Model Evaluation

After cross-validation and testing, both models performed strongly, but XGBoost consistently achieved higher Recall, Precision, ROC-AUC and Accuracy.

Metric	Random Forest	XGBoost
ROC-AUC	0.96	1.00
Accuracy	0.89	0.92
Precision (GBM)	0.87	1.00
Recall (GBM)	0.89	0.92

Interpretation:
Both models generalized well, with no major overfitting observed.
XGBoost captured subtler patterns in mutation interactions, improving recall of GBM, which is critical for clinical use.

6. Feature Importance

XGBoost revealed biologically coherent patterns:

Rank	Feature	Interpretation
1	IDH1_MUTATED	Hallmark of LGG.
2	PTEN_MUTATED	Classic GBM-associated tumor.
3	AGE 61–80	Older age group correlates with higher probability of GBM vs LGG.
4	Primary_Diagnosis_ Astrocytoma	Astrocytoma diagnosis aligns more with LGG profiles.
5	Primary_Diagnosis_Glioblastoma	Clinical label consistent with GBM.

The confusion matrix showed:

- Most misclassifications occurred in borderline cases (e.g. LGG with mixed genetic profiles).
- False negatives (GBM predicted as LGG) were minimal, which is primordial to avoid underestimating tumor severity.

Conclusions

The conducted study demonstrates that a data-driven approach using ensemble machine learning can reliably classify glioma grade, using clinical and genetic data.
The best performing model, XGBoost, achieved a Recall of 0.92, indicating high discriminatory power.
Key genetic markers (like IDH1, EGFR, PTEN, TP53) played dominant roles in distinguishing LGG from GBM, backed by well-known oncological evidence and reinforcing the model's biological interpretability.
This work validates the potential of AI-assisted clinical support systems in oncology, particularly for early detection and diagnostic triage. The results also highlight the importance of robust data preprocessing, proper validation and metric selection in medical ML applications.

Recommendation for managers

1. **Adopt the model as a decision-support tool** (not a diagnostic replacement): The model can assist clinicians in identifying probable GBM cases early, supporting prioritization and resource allocation.
2. **Prioritize recall for GBM cases**: In deployment, thresholds should favor sensitivity to avoid missing high-grade cases.
3. **Continuous monitoring and retraining**: Regular updates with new data are essential to maintain model performance as genomic knowledge evolves.
4. **Ensure explainability and regulatory compliance**: Incorporate SHAP or feature attribution tools before deployment in clinical stages to guarantee transparency and trustworthiness.
5. **Explore multimodal integration**: Combining genomic data with MRI imaging or histopathology could further enhance predictive accuracy.