Technical Annex – Glioma Grading Classification

Group 5 – Group Assignment, Machine Learning 2

1. Overview

This technical annex documents the detailed methodology, preprocessing, model training, and evaluation steps of Group 5's Glioma Grading Classification Project.

The objective was to classify glioma tumors into Low-Grade Glioma (LGG) and Glioblastoma Multiforme (GBM) using clinical, demographic, and genetic mutation data obtained from Kaggle (adapted from The Cancer Genome Atlas – TCGA).

The annex complements the main summary report by providing the technical foundation, including preprocessing details, parameter grids, evaluation metrics, and visual analyses.

2. Data Preparation

The dataset includes approximately 1,500 patient records and over 40 variables, covering three main domains:

- Clinical (e.g., Primary Diagnosis)
- Demographic (e.g., Age, Gender, Race)
- Genetic mutations (binary indicators for oncogenes and tumor suppressor genes)

2.1 Handling Missing Values

- Replaced placeholder values ('--', 'not reported') with 'Unknown' to preserve data integrity.
- Verified that the target column (Grade) contained no missing entries.

2.2 Age Transformation

- Converted text entries like "55 years 210 days" to numerical ages.
- Binned numeric ages into five intervals:
  0–20, 21–40, 41–60, 61–80, and 80+.
- Created a new variable Age_bin, later one-hot encoded.
- Unreported ages have been associated to the first bin.

2.3 Encoding

- Mutation features: converted from text ("MUTATED", "NOT_MUTATED") to numeric (1 / 0).
- Categorical variables: (Gender, Race, Primary_Diagnosis, Age_bin) one-hot encoded using OneHotEncoder(drop='first').
- Combined all encoded columns into a single dataframe of fully numeric features.

2.4 Feature Reduction and Redundancy Handling

To ensure interpretability and remove redundant or irrelevant variables, three filtering thresholds were applied:

1. High redundancy filter:
    - Features with redundancy ≥ 0.99 were discarded (mutations appearing in nearly all patients).
2. Low relevance filter:
    - Features with relevance ≤ 0.0001 were removed (mutations too rare to affect model learning).

After filtering, the final dataset retained 41 features, including clinical, demographic, and mutation variables, and was ready for modelling.

3. Modelling and Hyperparameter Tuning

Two ensemble learning algorithms were trained and compared:

Model	Principle	Key Strength
Random Forest	Bagging (Bootstrap Aggregation)	Stable, interpretable, resistant to overfitting
XGBoost	Gradient Boosting	Sequential error correction, high predictive power, efficient handling of class imbalance

3.1 Model Training Configuration

- Train/Test Split: 80/20
- Validation: 3-Fold Stratified Cross-Validation (CV)
- Scoring Metric: Recall, Precision, Accuracy, ROC-AUC
- Random Seed: 42 (ensures reproducibility)

3.2 Hyperparameter Grids

Random Forest:

n_estimators: [50, 100, 200]
max_depth: [3, 5, 7]
min_samples_split: [10, 20]
min_samples_leaf: [5, 10]
max_features: ['sqrt', 'log2']

XGBoost:

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]

The GridSearchCV procedure systematically tested these parameter combinations using Recall as the scoring metric.

4. Validation and Evaluation

A StratifiedKFold (k=3) CV design was used to maintain class balance between LGG and GBM in each fold.

The final models were retrained on the full training set using the best hyperparameters and evaluated on the 20% hold-out test set.

Performance Metrics

Metric	Random Forest	XGBoost	Interpretation	Clinical Relevance
ROC-AUC	0.96	1.00	Excellent separation ability	Reliable tumor grading
Accuracy	0.89	0.92	Strong overall generalization	Consistent patient-level predictions
Precision (GBM)	0.87	1.00	Very few false positives	Confidence in GBM detection
Recall (GBM)	0.89	0.92	Very few false negatives	Ensures aggressive tumors are flagged

The XGBoost model outperformed Random Forest on all metrics, especially in GBM recall, which is crucial to minimize false negatives in clinical practice.

5. Figures and Visual Analysis

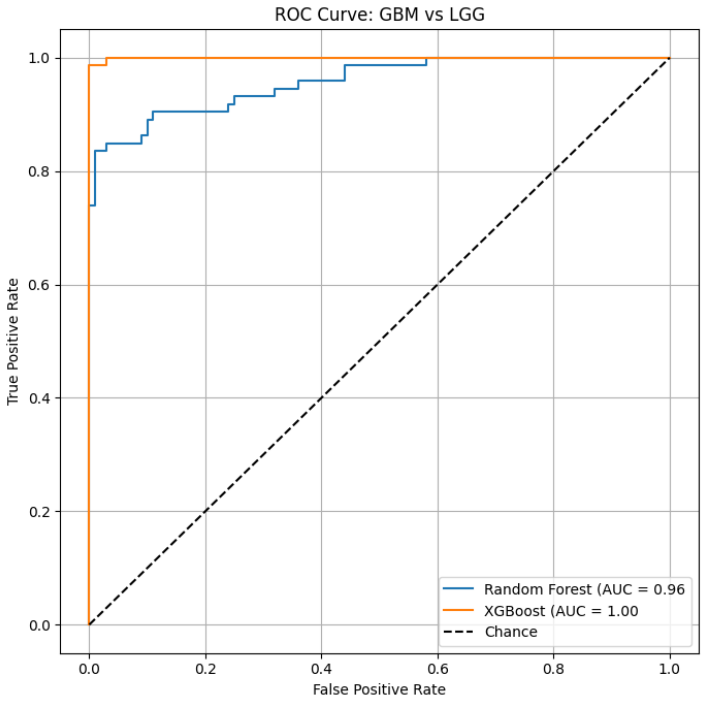

Figure 1: ROC Curves comparing Random Forest and XGBoost models.

Both models demonstrate AUC > 0.9, indicating strong separation between LGG and GBM.

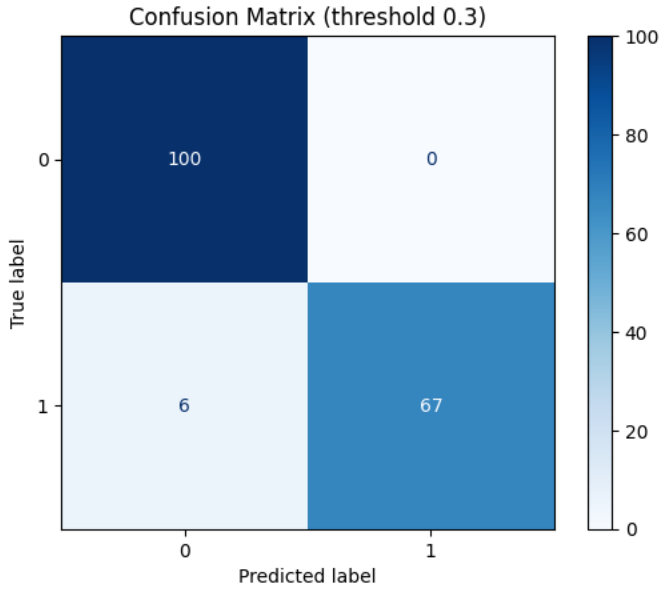

Figure 2: Confusion Matrix for XGBoost on test data.

GBM false negatives minimized, excellent recall performance.

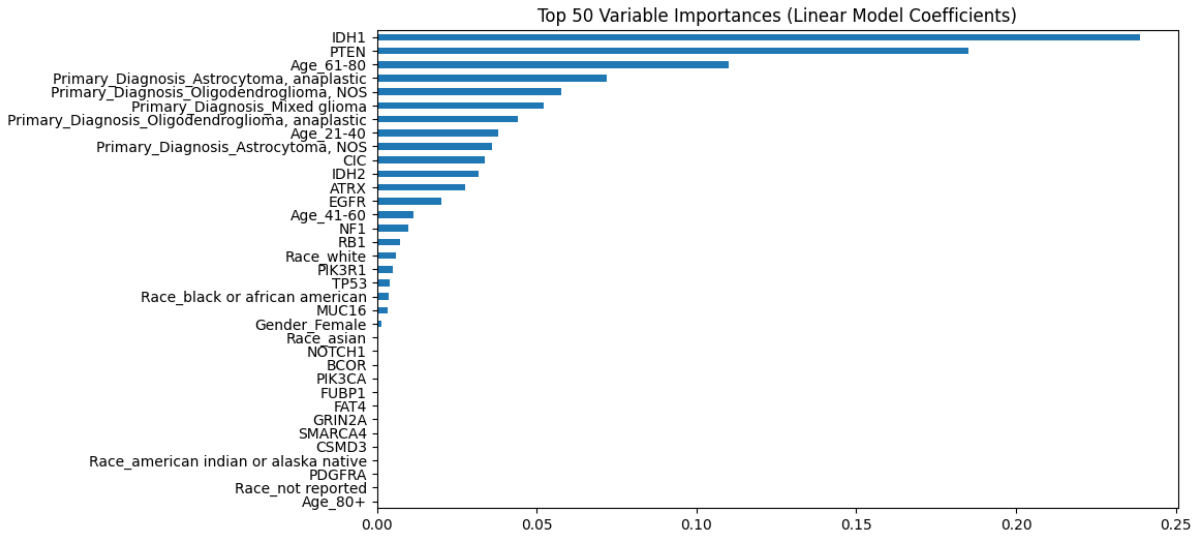

Figure 3: Feature Importance plot for XGBoost.

Top predictors: IDH1_MUTATED, PTEN_MUTATED, AGE 61-80, Primary_Diagnosis_ Astrocytoma**,** and Primary_Diagnosis_Glioblastoma.

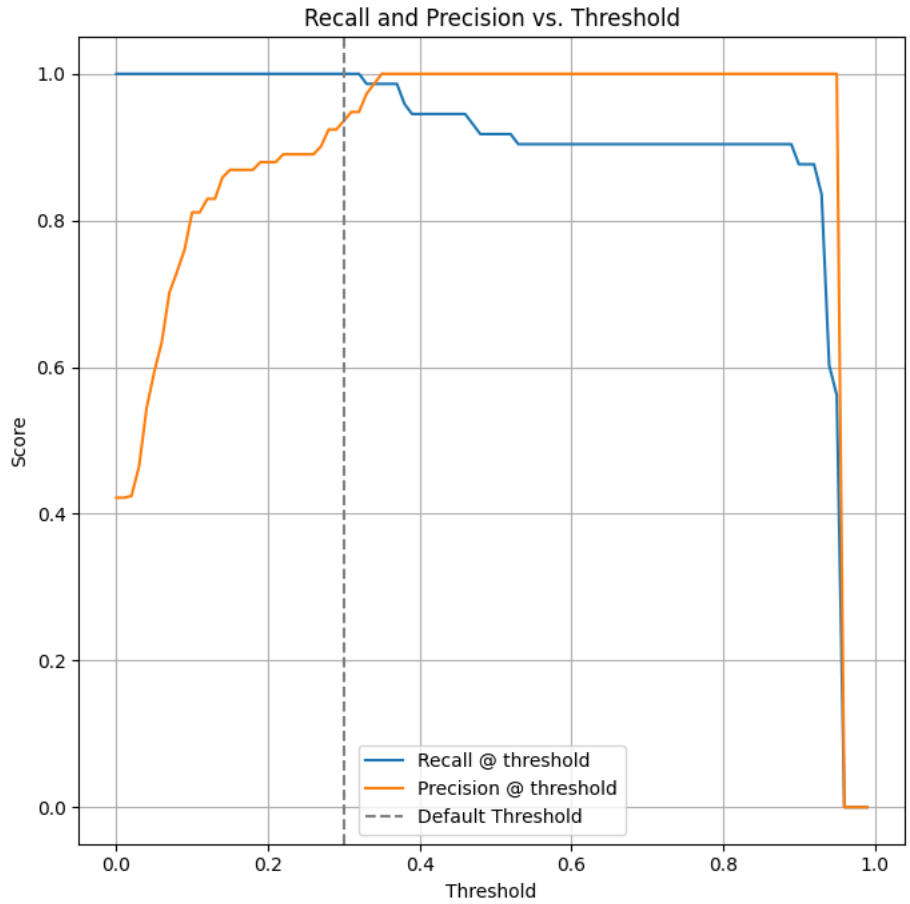

Figure 4: Precision–Recall trade-off curve.

Threshold near 0.3 favoring GBM recall to minimize false negatives.

6. Feature Importance and Biological Interpretability

Rank	Feature	Association	Interpretation
1	IDH1_MUTATED	↓ GBM	Hallmark of LGG; its presence strongly lowers GBM likelihood.
2	PTEN_MUTATED	↑ GBM	Classic GBM-associated tumor; increases GBM likelihood.
3	AGE 61–80	↑ GBM	Older age group correlates with higher probability of GBM vs LGG.
4	Primary_Diagnosis_ Astrocytoma	↓ GBM	Astrocytoma diagnosis aligns more with LGG profiles.
5	Primary_Diagnosis_Glioblastoma	↑ GBM	Clinical label consistent with GBM; strong positive association.

The top features identified by XGBoost align strongly with known oncological literature, confirming that the model captures genuine biological patterns rather than random statistical noise.

7. Technical Environment

Component	Description
Programming Language	Python 3.11
Key Libraries	pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost
Validation Tool	GridSearchCV (scikit-learn)
Cross-Validation	StratifiedKFold (k=3)
Random Seed	42 (ensures reproducibility)
Hardware	Local execution (8 GB RAM)

8. Limitations and Future Work

While the results demonstrate high recall, several limitations remain:

- The model is based solely on TCGA-derived data; additional validation with external hospital datasets is necessary.
- It does not include radiomic or imaging features, which may further enhance predictive precision.
- Although interpretable via feature importance, implementing SHAP explainability would provide case-level insights.
- Future work should integrate multi-modal inputs (genomic + imaging) and perform prospective validation in a clinical setting.

9. Conclusions

The technical process successfully produced a robust, interpretable model capable of classifying glioma grade with high reliability (Recall (GBM) = 0.92).

Through systematic preprocessing, feature filtering, and model tuning, the XGBoost classifier demonstrated superior predictive performance and strong alignment with known biological mechanisms.

This annex validates the methodology and technical soundness of the project, complementing the summary report's managerial and clinical interpretation.