Classifying Glioma Grade

Predicting Glioma Severity Using Machine Learning: A Data-Driven Clinical Support Tool

Group 5

Paula Caceres
Christelle El-Haddad
Radwan Fenaer
Regina Ortega
Flavio Valerio
Luis Wilhelmi

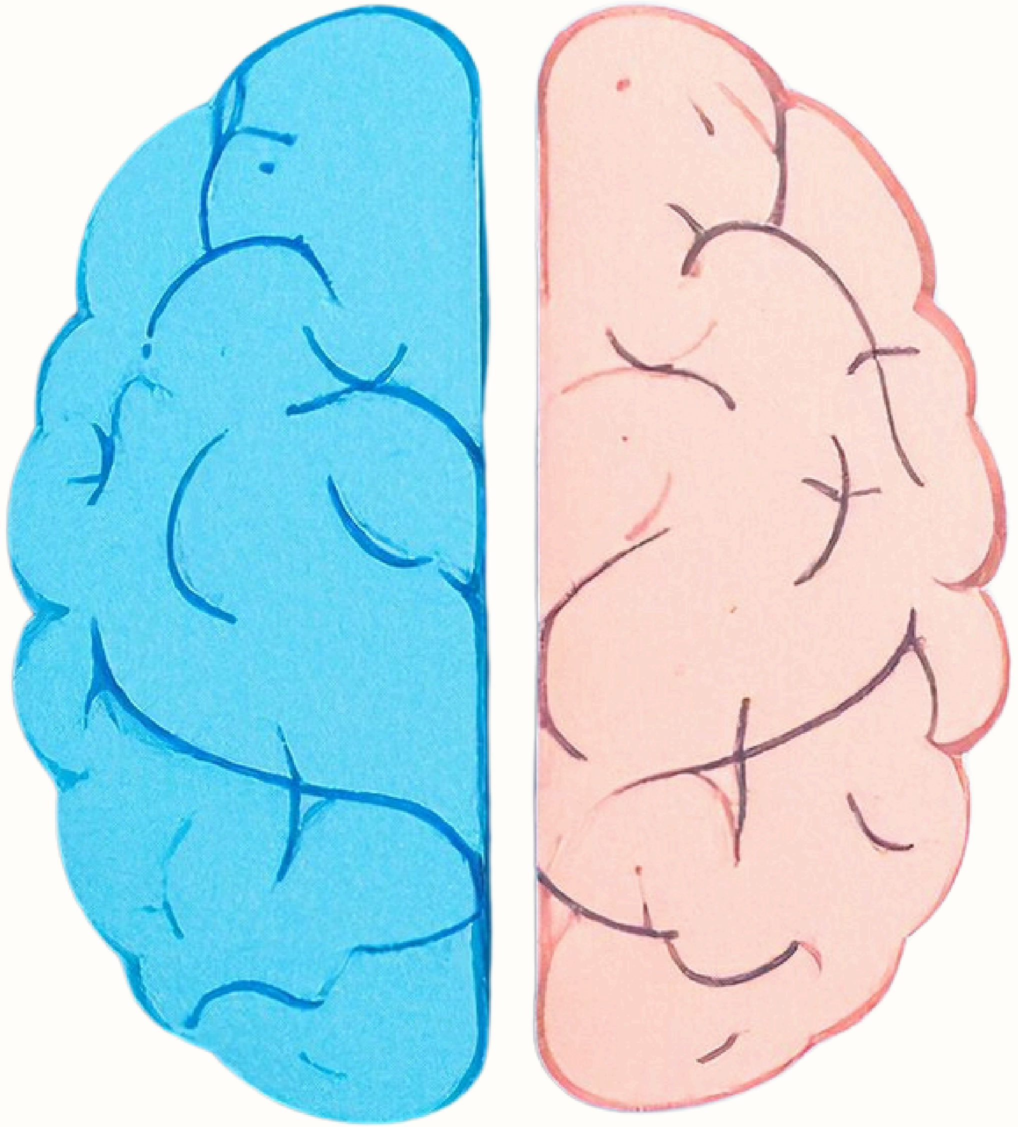
What’s Glioblastoma

- 1 in 100 cancers affects the brain.
- Gliomas: most common group.
- Glioblastoma: very aggressive type of gliomas.

What if we could predict which tumors are the most dangerous, faster?

Gliomas Classification

Low-Grade Glioma (LGG)
→ Slow-growing



Glioblastoma Multiforme (GBM)
→ Aggressive

Two glioma subtypes - radically different prognoses

Challenge

Early glioma grade classification and problem motivation

Methodology

Dataset Overview,
Modelling steps,
Model comparaison

The Model

Insights on model performance and biological interpretation

Conclusion

Implications, recommendations, and next steps

Challenge

Can a machine learning model help identify whether a brain tumor is likely to be low-grade or high-grade?

Why Predicting Glioma Severity Matters

Patient Impact

- GBM is aggressive - early grading saves lives.
- Vastly different treatments.

Operational impact

- Current diagnostics: invasive and time-consuming.
- Histopathology or MRI imaging.

Opportunity

- Machine learning classifier predicting glioma grade from clinical and genomic data to support early triage and clinicians.

Key Questions

1. Can we accurately predict glioma grade (LGG vs GBM) using non-imaging data?

2. Which clinical and genetic factors drive this prediction?

3. Which machine learning model provides the best balance between performance and interpretability?

Methodology

A step-by-step process to understand our data driven modelling approach

Dataset Overview

Integrating clinical, demographic, and genetic data for glioma classification

- **Source**

  From kaggle, adapted from The Cancer Genome Atlas (TCGA) project.

- **Samples**

  ~1,500 patients with primary brain tumors.

- **Features**

  30+ clinical, demographic, and gene-mutation variables.

- **Target variable**

  “Grade”: **0 = LGG, 1 = GBM.**

- **Key data domains**

  - **Demographic** (age, gender, race, etc.).
  - **Clinical** (primary diagnosis, etc.).
  - **Genetic** (mutation types in binary format).

Key Variables & Data Domains

Clinical and genetic features that drive grade differences

Mutational landscape

- **IDH1 mutation**: hallmark of LGG (~80%).
- **EGFR/PTEN mutations**: markers for GBM.
- **TP53 / ATRX mutations**: shared across both, useful interaction terms.

Age

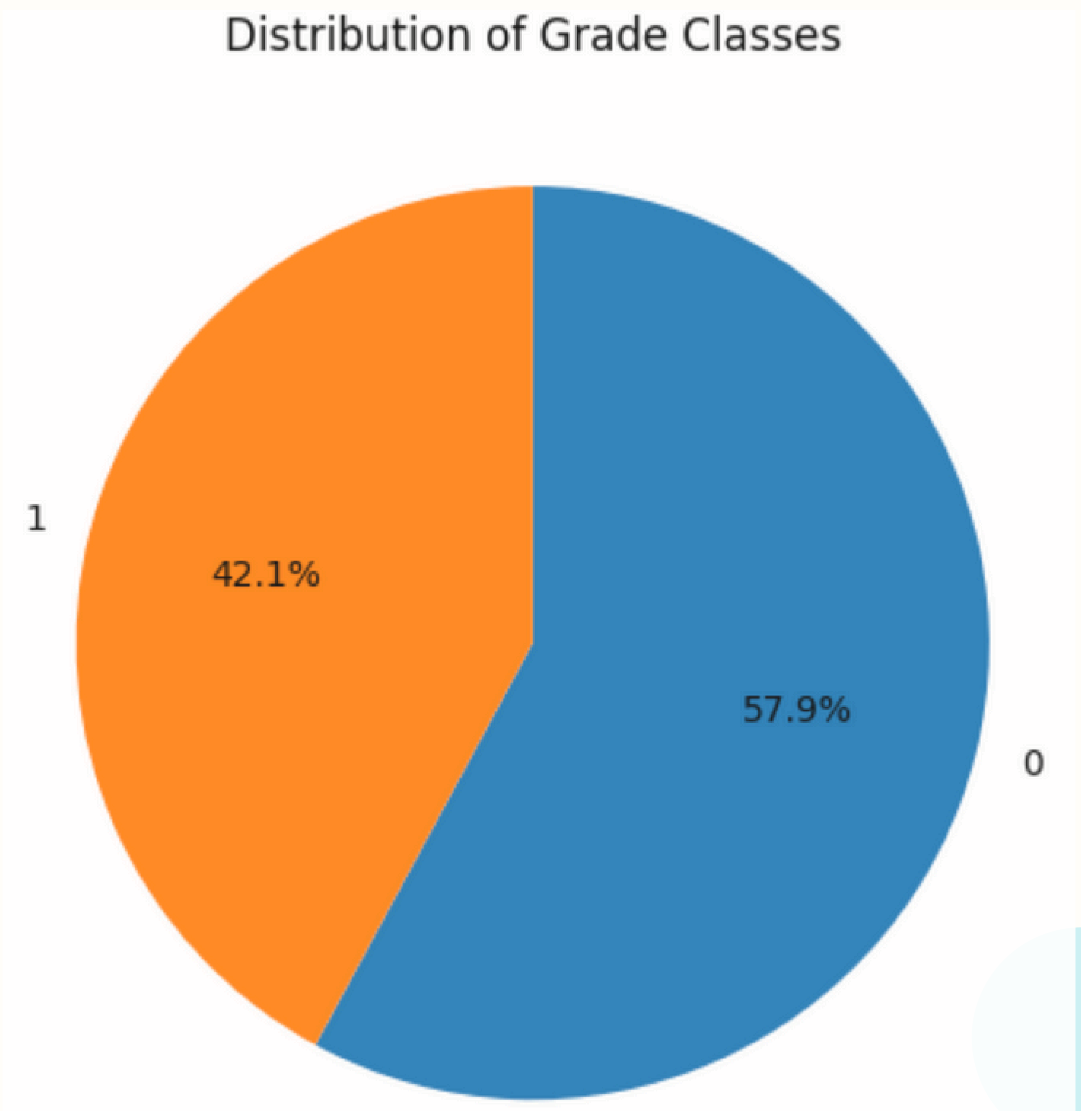
GBM patients are typically older (mean 59) than LGG patients (mean 46).

Primary diagnosis

“Glioblastoma” strongly associated with GBM, “Astrocytoma” & “Oligodendroglioma” mostly to LGG.

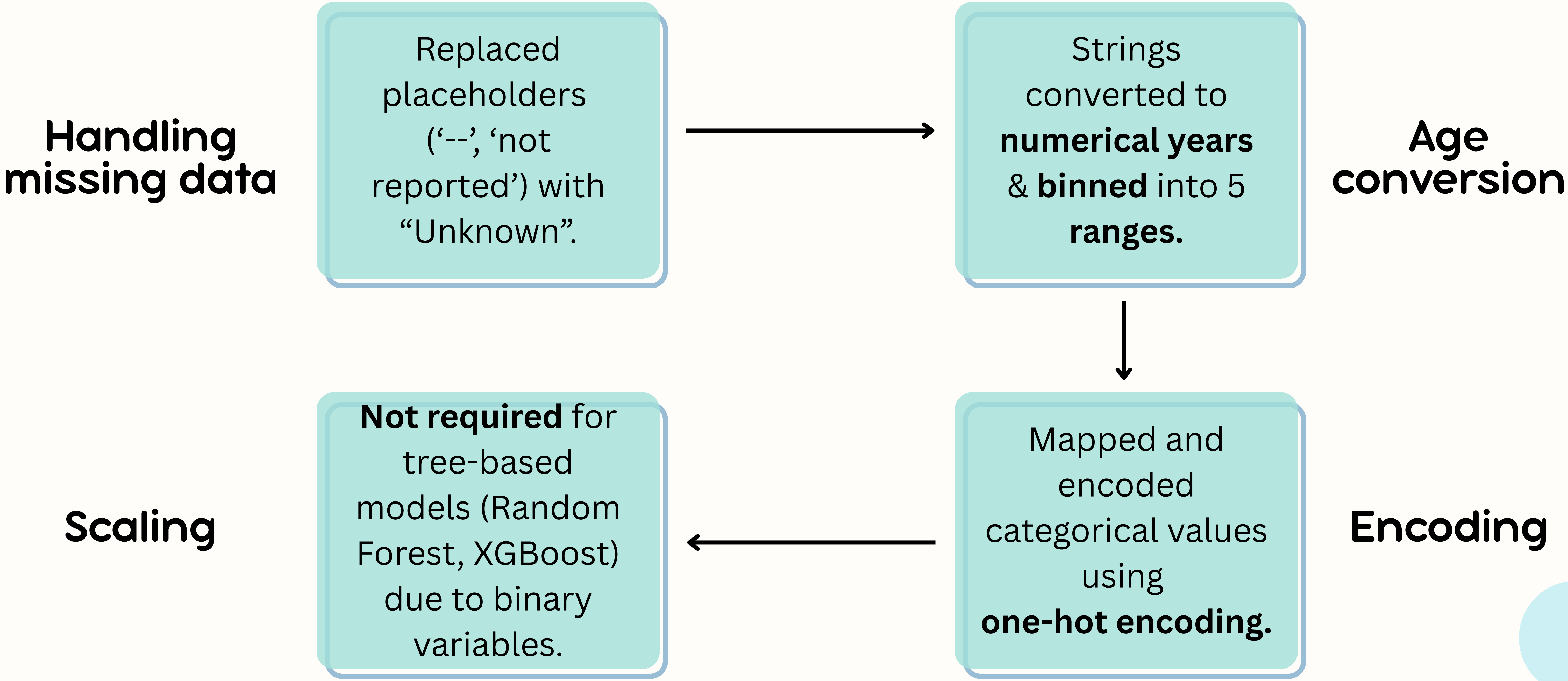
Clinical alignment

Matches established oncology findings, validating dataset quality.



Data Cleaning & Encoding

Transforming raw metadata into machine-readable features



Handling missing data

Replaced placeholders (‘--’, ‘not reported’) with “Unknown”.

Age conversion

Strings converted to **numerical years** & **binned** into 5 **ranges.**

Encoding

Mapped and encoded categorical values using **one-hot encoding.**

Scaling

Not required for tree-based models (Random Forest, XGBoost) due to binary variables.

Feature Reduction & Correlation Check

Preventing multicollinearity and information leakage

- **Approach**

  Pearson correlation analysis among numeric variables.

- **Threshold**

  Removed features with redundancy > 0.99 and irrelevance < 0.0001.

- **Outcome**

  Reduced from 43 to 41 features, keeping all major biological variables.

Modeling Strategy

Comparing ensemble models for tabular genomic data

Algorithms tested

- **Random Forest (Bagging)** - stable, interpretable, strong baseline.
- **XGBoost (Boosting)** - sequential, captures complex relationships.

Why ensemble methods

- Handle mixed feature types (numeric + categorical).
- Robust to scaling and missing data.
- Offer feature importance for interpretability.

Evaluation metric

- **Recall** - preferred for critical nature of clinical cases.

Validation & Hyperparameter Tuning

Ensuring robustness and generalization

- **Validation Design**
  - 3-fold **Stratified Cross-Validation (CV)** to preserve GBM/LGG ratio.
  - Split into Train (80%) / Test (20%).

- **Optimization**

  GridSearchCV explored parameter grids:
  - **Random Forest**: n_estimators, max_depth, min_samples_split.
  - **XGBoost**: learning_rate, max_depth, n_estimators.

- **Metric**

  Recall on validation folds, best model generalized.

THE MODEL

Evaluating model performance, interpretability, and clinical relevance

Model Performance During Cross-Validation

Comparing Random Forest and XGBoost on validation folds

- **Random Forest**

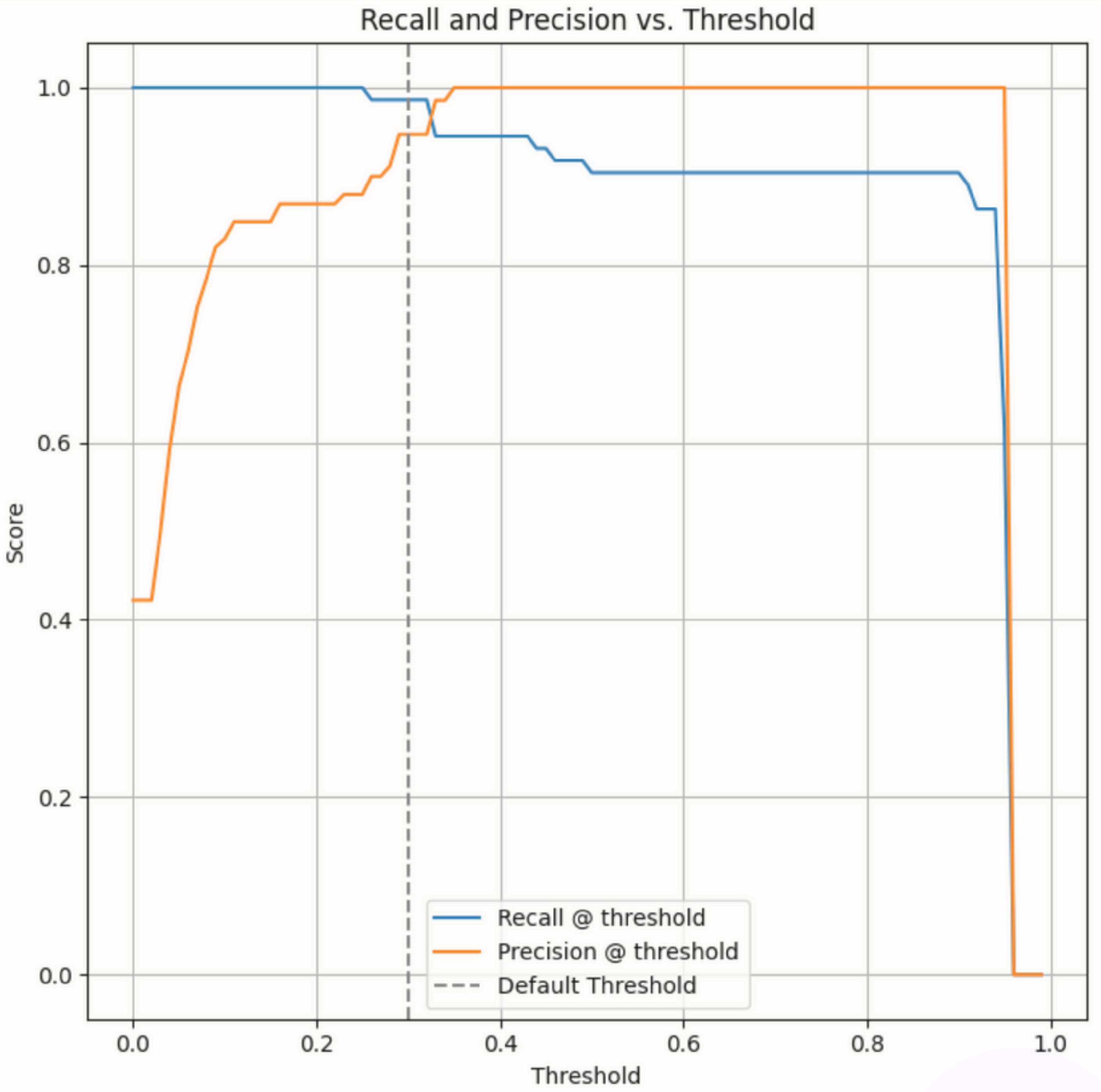
  Test Recall = 0.89
  Test Precision = 0.87

- **XGBoost**

  Test Recall = 0.92
  Test Precision = 1.00

- **Main differentiator**

  Both capture class separation well, but XGBoost handles non-linear feature interactions better.

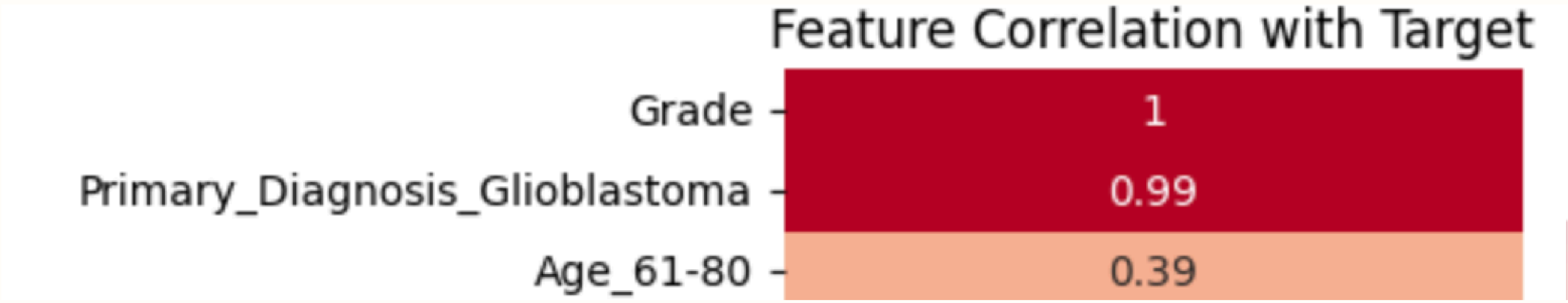


Feature Importance and Interpretability

Top 3 most important variables for the model

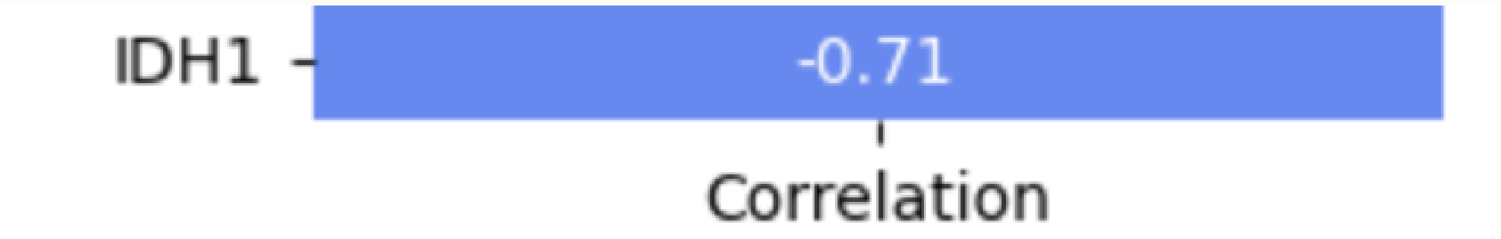
Positive correlation

Positively correlated features highlight the key clinical and genetic drivers of tumor aggressiveness.



Negative correlation

Negatively correlated features represent protective or low-grade genetic and clinical patterns.



Test Performance and Evaluation Metrics

Evaluating accuracy, recall, and precision for GBM classification

- **Interpretation**

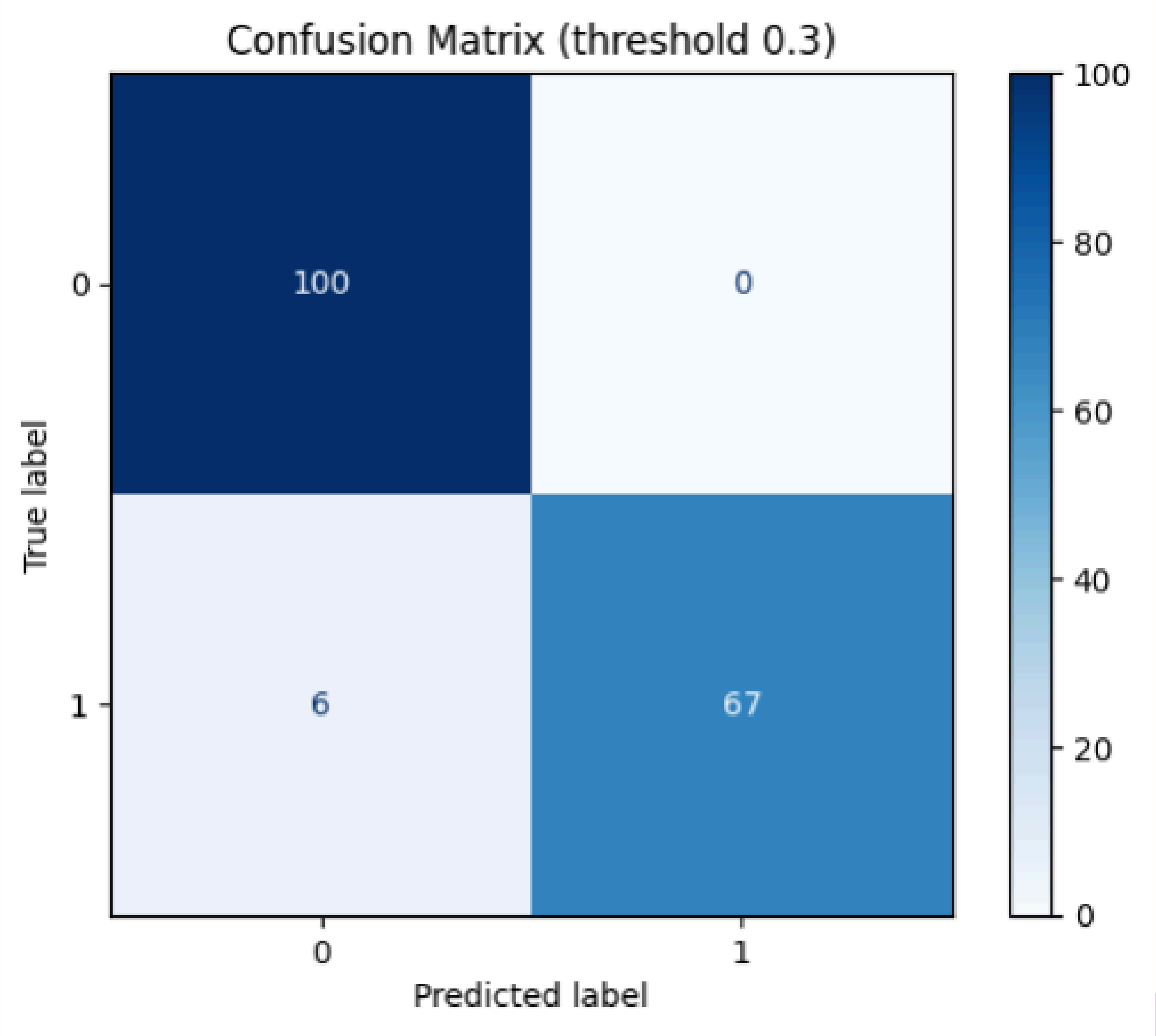
  XGBoost achieves the best balance of precision and recall.

- **Clinical focus**

  High recall for GBM, fewer missed aggressive tumors (FN).

- **Model confidence**

  Recall > 0.9 = high distinction between LGG & GBM.



Error Analysis

Comparing Random Forest and XGBoost on validation folds

- **Borderline profiles**

  LGG samples with rare GBM-like mutations (e.g., EGFR, TP53) misclassified as GBM.

- **Few GBM false negatives**

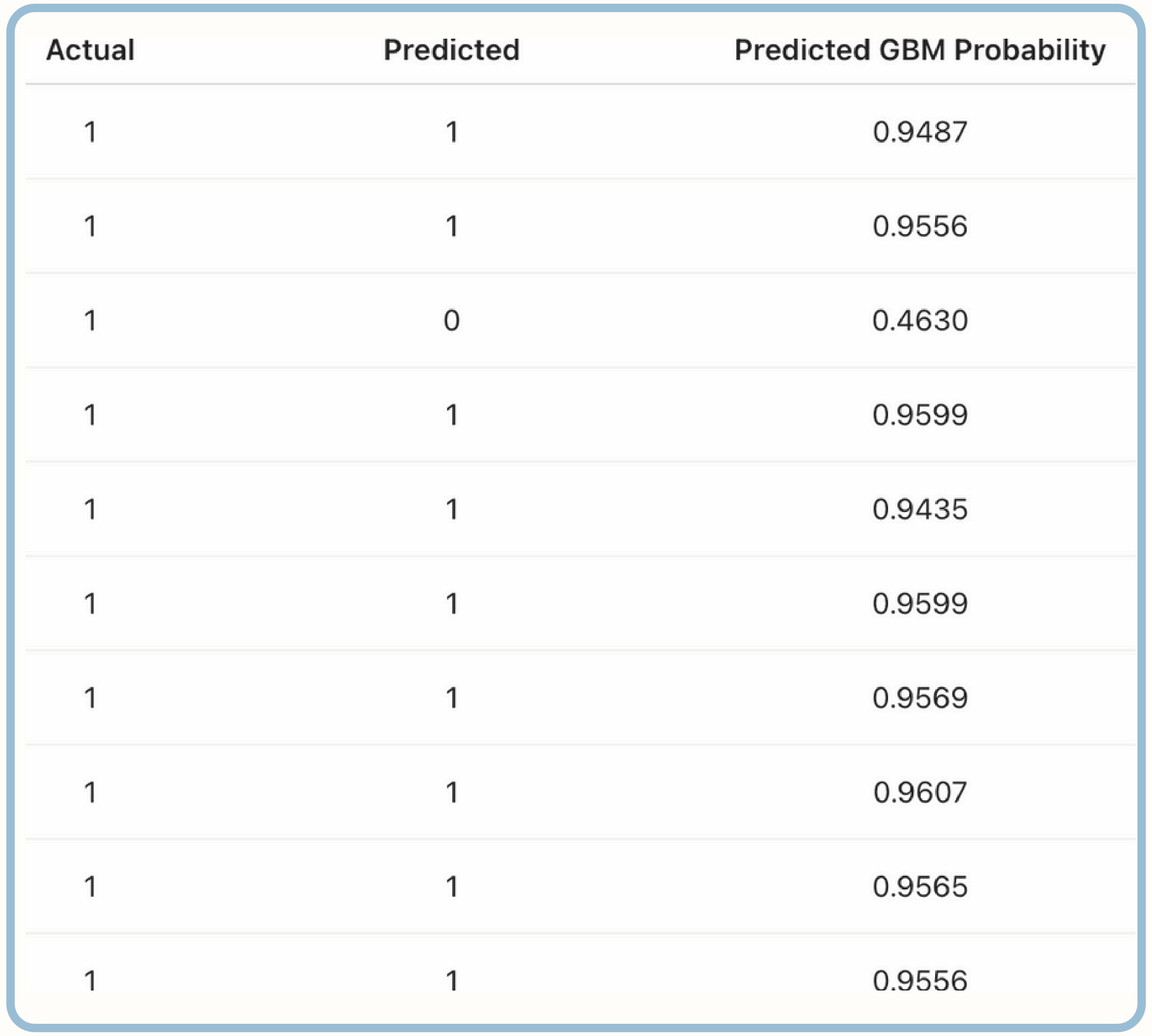
  Critical goal met = low clinical risk.

- **Random Forest**

  Higher variance in borderline cases.

- **XGBoost**

  Better at capturing non-linear mutation interactions → fewer errors.



Actual	Predicted	Predicted GBM Probability
1	1	0.9487
1	1	0.9556
1	0	0.4630
1	1	0.9599
1	1	0.9435
1	1	0.9599
1	1	0.9569
1	1	0.9607
1	1	0.9565
1	1	0.9556

Actual vs. predicted values for the last 10 rows

Conclusion

From predictive performance to real-world medical impact

Conclusion

From model accuracy to actionable clinical value

Project Achievement

- Built a machine learning pipeline to classify glioma grade (LGG vs GBM).
- **XGBoost** achieved the best performance (**Recall 0.92**).

Broader Impact

- Supports **faster, data-driven diagnosis** and early triage.
- A step toward **AI-assisted precision oncology**.

Core Insights

- Top features (**IDH1, PTEN, Age 61-80, Astrocytoma, Glioblastoma**) match clinical biology.

Recommendations

Potential next steps

1. Adopt the model as a decision-support tool (not a diagnostic replacement).

2. Prioritize recall for GBM cases.

3. Continuous monitoring and retraining.

4. Ensure explainability and regulatory compliance.

5. Explore multimodal integration.

Thank you